General bounds on the quality of Bayesian coresets

Trevor Campbell*
Department of Statistics
University of British Columbia
trevor@stat.ubc.ca

Abstract

Bayesian coresets speed up posterior inference in the large-scale data regime by approximating the full-data log-likelihood function with a surrogate log-likelihood based on a small, weighted subset of the data. But while Bayesian coresets and methods for construction are applicable in a wide range of models, existing theoretical analysis of the posterior inferential error incurred by coreset approximations only apply in restrictive settings—i.e., exponential family models, or models with strong log-concavity and smoothness assumptions. This work presents general upper and lower bounds on the Kullback-Leibler (KL) divergence of coreset approximations that reflect the full range of applicability of Bayesian coresets. The lower bounds require only mild model assumptions typical of Bayesian asymptotic analyses, while the upper bounds require the log-likelihood functions to satisfy a generalized subexponentiality criterion that is weaker than conditions used in earlier work. The lower bounds are applied to obtain fundamental limitations on the quality of coreset approximations, and to provide a theoretical explanation for the previously-observed poor empirical performance of importance sampling-based construction methods. The upper bounds are used to analyze the performance of recent subsample-optimize methods. The flexibility of the theory is demonstrated in validation experiments involving multimodal, unidentifiable, heavy-tailed Bayesian posterior distributions.

1 Introduction

Large-scale data is now commonplace in scientific and commerical applications of Bayesian statistics. But despite its prevalence, and the corresponding wealth of research dedicated to scalable Bayesian inference, there are still suprisingly few general methods that provably provide inferential results, within some reasonable tolerated error, at a significant computational cost savings. Exact Markov chain Monte Carlo (MCMC) methods require many full passes over the data [1, Ch. 6–12, 2, Ch. 11–12], limiting the utility of these methods when even a single pass is expensive. A wide range of MCMC methods that access only a subset of data per iteration, e.g., via delayed acceptance [3–6], pseudomarginal or auxiliary variable methods [7–9], and basic subsampling [10–13], provide at most a minor improvement over full-data MCMC [14–16]. On the other hand, methods including carefully constructed log-likelihood function control variates can provide substantial gains [17–19]. However, black-box control variate constructions for large-scale data often rely on assumptions such as posterior density differentiability and unimodality that do not hold in many popular models, e.g., those with discrete variables or multimodality. See [15, 20] for a survey of scalable MCMC methods. Parametric approximations via variational inference [21] or the Laplace approximation [22, 23] can be obtained scalably using stochastic optimization methods, but existing general theoretical guarantees for these methods again typically rely on posterior normality assumptions [24, p. 141–144,25–30] (see [21, 31] for a review).

*https://trevorcampbell.me

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

Although many existing methods rely on asymptotic normality or unimodality in the large-scale data regime, the problem of handling large-scale data in Bayesian inference does not fundamentally require this structure. Instead, one can more generally exploit *redundancy* in the data (i.e., the existence of good approximate sufficient statistics), which can be used to draw principled conclusions about a large data set based only on a small fraction of examples. Indeed, while approximate posterior normality often does not hold in models with latent discrete or combinatorial objects, weakly identifiable or unidentifiable parameters, persisting heavy tails, multimodality, etc., such models can and regularly do exhibit significant redundancy in the data that can be exploited for faster large-scale inference. *Bayesian coresets* [32]—which involve replacing the full dataset during inference with a sparse weighted subset—are based on this notion of exploiting data redundancy. Empirical studies have shown the existence of high-quality coreset posterior approximations constructed from a small fraction of the data, even in models that violate posterior normality assumptions and for which standard control variate techniques work poorly [33–37]. However, existing theoretical support for Bayesian coresets in the literature is limited. There exist no lower bounds on Bayesian coreset approximation error, and while upper bounds do exist, they currently impose restrictive assumptions. In particular, the best available theoretical upper bounds to date apply to exponential family models [36, 38] and models with strongly log-concave and locally smooth log-densities [37].

This article presents new theoretical techniques and results regarding the quality of Bayesian coreset approximations. The main results are two general large-data asymptotic lower bounds on the KL divergence (Theorems 3.3 and 3.5), as well as a general upper bound on the KL divergence (Theorem 5.3) under the assumption that the log-likelihoods satisfy a multivariate generalization of subexponentiality (Definition 5.2). The main general results in this paper lead to various novel insights about specific Bayesian coreset construction methods. Under mild assumptions,

- common importance-weighted coreset constructions (e.g. [32]) require a coreset size M proportional to the dataset size N (Corollary 4.1), even with post-hoc optimal weight scaling (Corollary 4.2), and thus yield a negligible improvement over full-data inference;
- *any* construction algorithm requires a coreset size $M > d$ when the log-likelihood function is determined by d parameters locally around a point of concentration (Corollary 4.3);
- subsample-optimize coreset construction algorithms (e.g. [36–39]) achieve an asymptotically bounded error with a coreset size $\mathrm{polylog} N$ in a wide variety of models (Corollary 6.1).

The paper includes empirical validation of the main theoretical claims on two models that violate common assumptions made in the literature: a multimodal, unidentifiable Cauchy location model with a heavy-tailed prior, and an unidentifiable logistic regression model with a heavy-tailed prior and persisting posterior heavy tails. Experiments were performed on a computer with an Intel Core i7-8700K and 32GB of RAM. Proofs of all theoretical results may be found in Appendix A.

Notation. We use standard asymptotic growth symbols $O, \Omega, \Theta, o, \omega$ (see, e.g., [40, Sec. 3.3]), and their probabilistic variants $O_p, \Omega_p, \Theta_p, o_p, \omega_p$ (see, e.g., [24, Sec. 2.2]). We use the same symbol to denote a measure π and its density $\pi(\cdot)$ with respect to a specified dominating measure. We also regularly suppress integration variables and differential symbols in integrals throughout for notational brevity when these are clear from context; for example, $\int \pi \exp(\ell)$ is shorthand for $\int \pi(\mathrm{d}\theta) \exp(\ell(\theta))$. Finally, the pushforward of a measure π by a map η is denoted simply $\eta\pi$.

2 Background

Define a target probability distribution π on a space Θ comprised of a sum of N potentials $\ell_n : \Theta \to \mathbb{R}$, $n = 1, \ldots, N$ and a base distribution $\pi_0(\mathrm{d}\theta)$,

$$\pi(\mathrm{d}\theta) = \frac{1}{Z} \exp\left(\ell(\theta)\right) \pi_0(\mathrm{d}\theta), \qquad \ell(\theta) = \sum_{n=1}^{N} \ell_n(\theta), \qquad \theta \in \Theta,$$

where the normalization constant Z is not known. In the Bayesian context, this distribution corresponds to a Bayesian posterior distribution for a statistical model with prior π_0 and conditionally i.i.d. data X_n, where $\ell_n(\theta) = \log p(X_n|\theta)$. The goal is to compute or approximate expectations under π; but the likelihood ℓ (and its gradient) becomes expensive to evaluate when N is large. To

avoid this cost, *Bayesian coresets* [32–37] involve replacing the target with a surrogate density

$$\pi_w(\mathrm{d}\theta) = \frac{1}{Z(w)} \exp\left(\ell_w(\theta)\right) \pi_0(\mathrm{d}\theta), \qquad \ell_w(\theta) = \sum_{n=1}^{N} w_n \ell_n(\theta), \qquad \theta \in \Theta,$$

where $w \in \mathbb{R}^N$, $w \geq 0$ are a set of weights, and $Z(w)$ is the new normalizing constant. If w has at most $M \ll N$ nonzeros, the $O(M)$ cost of evaluating $\sum_n w_n \ell_n$ (and its gradient) is a significant improvement upon the original $O(N)$ cost. In this work, the problem of coreset construction is formulated in the data-asymptotic limit; a coreset construction method should

- run in $o(N)$ time and memory (or at most $O(N)$ with a small leading constant),
- produce a small coreset of size $M = o(N)$,
- produce a coreset with $O(1)$ posterior forward/reverse KL divergence as $N \to \infty$.

These three desiderata ensure that the effort spent constructing and sampling from the coreset posterior is worthwhile: the coreset provides a meaningful reduction in computational cost compared with standard Markov chain Monte Carlo algorithms, and has a bounded approximation error.

3 Lower bounds on approximation error

This section presents lower bounds on the KL divergence of coreset approximations for general models and data generating processes. The first key steps in the analysis are to write all expectations in terms of distributions that do not depend on w, and to remove the difficult-to-control influence of the tails of π and π_w by restricting certain integrals to some small subset $B \subseteq \Theta$ of the parameter space. Lemma 3.1, the key theoretical tool used in this section, achieves both of these two goals; note that the result has no major assumptions and applies generally in any setting that a Bayesian coreset can be used. For convenience, define

$$\underline{\mathrm{KL}}(w) := \min\{\mathrm{KL}(\pi_w || \pi), \mathrm{KL}(\pi || \pi_w)\},$$

and the decreasing, nonnegative function $f : \mathbb{R}_+ \to \mathbb{R}_+$,

$$f(x) = \begin{cases} -\log x + x - 1 & 0 \leq x \leq 1 \\ 0 & x > 1. \end{cases}$$

Lemma 3.1 (Basic KL Lower Bound)**.** *For all measurable $B \subseteq \Theta$ and coreset weights w,*

$$\underline{\mathrm{KL}}(w) \geq f(J_B(w)) \geq 0,$$

where

$$J_B(w) = \frac{\int_B \pi_0 \exp \frac{1}{2}(\ell + \ell_w)}{\sqrt{\int \pi_0 \exp(\ell) \int \pi_0 \exp(\ell_w)}} + \sqrt{\pi(B^c)}.$$

Note that while the integrals in the fraction denominator in $J_B(w)$ range over the whole Θ space, a further lower bound on $\underline{\mathrm{KL}}(w)$ can be obtained by restricting their domains arbitrarily. Also, crucially, the bound in Lemma 3.1 does not depend on $\pi_w(B^c)$, which would be difficult to analyze without detailed knowledge of the tail behaviour of π_w as a function of the coreset weights w. Although the bound in Lemma 3.1 applies generally, it is most useful when B is small (so that simple local approximations of ℓ and ℓ_w can be used), π concentrates on B (so that $\pi(B^c) \approx 0$), and π and π_w are very different when restricted to B; the behaviour of the bound in this case is roughly (see the proof in Appendix A) $f(J_B(w)) \approx -\log(1 - \mathrm{TV}(\pi, \pi_w))$. Finally, note that Lemma 3.1 remains valid if one replaces ℓ_w with $\ell_w - c$ and ℓ with $\ell - c'$ for any constants c, c' that do not depend on θ but may depend on the data and coreset weights w.

For the remainder of this section, consider the setting where Θ is a measurable subset of $\mathbb{R}^d$ for some $d \in \mathbb{N}$, fix some $\theta_0 \in \Theta$, and assume each ℓ_n is differentiable in a neighbourhood of θ_0. Let

$$\overline{w} = \sum_n w_n \qquad g = \nabla \ell(\theta_0) \qquad g_w = \nabla \ell_w(\theta_0).$$

Theorems 3.3 and 3.5 characterize KL divergence lower bounds in terms of the sum of the coreset weights $\overline{w}$ and the log-likelihood gradients g, g_w. Intuitively for the full data set where all $w_n = 1$ and $\overline{w} = N$, and an i.i.d. data generating process from the likelihood with parameter θ_0, the central limit theorem asserts under mild conditions that $g_w/\overline{w} \xrightarrow{p} 0$ at a rate of $N^{-1/2}$. Theorems 3.3 and 3.5 below provide KL lower bounds when the coreset construction algorithm does not match this behavior. In particular, Theorem 3.3 provides results that are useful when $g_w/\overline{w} \xrightarrow{p} 0$ occurs reasonably quickly but slower than $N^{-1/2}$, while Theorem 3.5 strengthens the conclusion when $g_w/\overline{w} \xrightarrow{p} 0$ very slowly or not at all. The major benefit of Theorems 3.3 and 3.5 for analyzing coreset construction methods is that they reduce the problem of analyzing posterior KL divergence to the much easier problem of analyzing the 2-norm $\|\cdot\|_2$ of a weighted sum of random vectors in $\mathbb{R}^d$.

Consider a sequence $r \to 0$ as $N \to \infty$, and for a fixed matrix $H \succ 0$ let

$$B = \{\theta : (\theta - \theta_0)^T H (\theta - \theta_0) \leq r^2\}$$

be a sequence of neighbourhoods around θ_0; these will appear in Assumptions 3.2 and 3.4 and Theorems 3.3 and 3.5 below. Note that throughout, all asymptotics will be taken as $N \to \infty$, and various sequences (e.g., r and B) are implicitly indexed by N. To simplify notation, this dependence is suppressed. Assumption 3.2 makes some weak assumptions about the model and data generating process: it intuitively asserts that the potential functions are sufficiently smooth around θ_0, that $r \to 0$ slowly, and that π concentrates at θ_0 at a usual rate. Note that Assumption 3.2 does not assume data are generated i.i.d. and places no conditions on the coreset construction algorithm.

Assumption 3.2. *π_0 has a density with respect to the Lebesgue measure, $\pi_0(\theta_0) > 0$, each $\ell_n(\theta)$ and $\pi_0(\theta)$ are twice differentiable in B for sufficiently large N, and*

$$\sup_{\theta \in B} \left\| -\frac{1}{N}\nabla^2 \ell(\theta) - H \right\|_2 = o_p(1), \quad \left\| \frac{g}{N} \right\|_2 = O_p\left(N^{-1/2}\right), \qquad Nr^2 = \omega(1).$$

Two additional assumptions related to the coreset construction algorithm—namely, that it works well enough that $\frac{1}{\overline{w}} \sum_n w_n \nabla^2 \ell_n(\theta) \xrightarrow{p} H$ and $g_w/\overline{w} \xrightarrow{p} 0$ at a rate faster than $r \to 0$—lead to asymptotic lower bounds on the best possible quality of coresets produced by the algorithm, as well as lower bounds even after optimal post-hoc scaling of the weights.

Theorem 3.3. *Suppose Assumption 3.2 holds. If*

$$\sup_{\theta \in B} \left\| -\frac{1}{\overline{w}}\nabla^2 \ell_w(\theta) - H \right\|_2 = o_p(1), \quad \left\| \frac{g_w}{\overline{w}} \right\|_2 = o_p(r),$$

then

$$\underline{\mathrm{KL}}(w) \geq O_p(1) + \Omega_p(1) \min\left\{ -\log \pi(B^c), \frac{N\overline{w}}{N+\overline{w}} \left\| \frac{g}{N} - \frac{g_w}{\overline{w}} \right\|_2^2 + d \log \frac{(N+\overline{w})^2}{N \max\{\overline{w}, 1/r^2\}} \right\}$$

$$\min_{\alpha \geq 0} \underline{\mathrm{KL}}(\alpha w) \geq O_p(1) + \Omega_p(1) \min\left\{ -\log \pi(B^c), d \log\left(N \left\| \frac{g}{N} - \frac{g_w}{\overline{w}} \right\|_2^2 \right) \right\}.$$

Theorem 3.3 is restricted to the case where the coreset algorithm is performing reasonably well. Theorem 3.5 extends the bounds to the case where the algorithm is performing poorly, in the sense that it is unable to make $\frac{g_w}{\overline{w}} \xrightarrow{p} 0$ at a rate faster than $r \to 0$ (or perhaps $\frac{g_w}{\overline{w}}$ does not converge to 0 at all). In order to draw conclusions in this setting, we need a weak global assumption on the potential functions. A function $f : \Theta \to \mathbb{R}$ is *L-smooth below at θ_0* if

$$\forall \theta \in \Theta, \quad f(\theta) \geq f(\theta_0) + \nabla f(\theta_0)^T (\theta - \theta_0) - \frac{L}{2} \|\theta - \theta_0\|_2^2. \tag{1}$$

Note that L-smoothness below is weaker than Lipschitz smoothness and does not imply concavity; Eq. (1) restricts the growth of the function only in the negative direction, and only when the expansion is taken at θ_0. Assumption 3.4 asserts that the potential functions are smooth below.

Assumption 3.4. *There exist $L_0, \ldots, L_N, L > 0$ such that $\log \pi_0$ is L_0^2-smooth below at θ_0, for each $n \in [N]$ ℓ_n is L_n^2-smooth below at θ_0, and $\frac{1}{N}\sum_{n=1}^N L_n^2 \xrightarrow{p} L^2$.*

Theorem 3.5 uses Assumptions 3.2 and 3.4 and additional assumptions related to the coreset construction algorithm to obtain lower bounds in a setting that relaxes the "performance" conditions in Theorem 3.3: $-\frac{1}{\overline{w}}\sum_n w_n \nabla^2 \ell_n(\theta)$ no longer needs to converge to H in probability, and $g_w/\overline{w}$ can converge to 0 slowly or not at all.

Theorem 3.5. *Suppose Assumptions 3.2 and 3.4 hold. If there exist $\alpha, \beta > 0$ such that*

$$\mathbb{P}\left(\forall \theta \in B,\ -\frac{1}{\overline{w}}\nabla^2 \ell_w(\theta) \succeq \alpha H\right) \to 1, \quad \mathbb{P}\left(\frac{1}{\overline{w}}\sum_n w_n L_n^2 \leq \beta L^2\right) \to 1, \quad \left\|\frac{g_w}{\overline{w}}\right\| = \omega_p(r),$$

then

$$\underline{\mathrm{KL}}(w) \geq O_p(1) + \Omega_p(1)\min\left\{-\log \pi(B^c), d\log\left(N \min\left\{\left\|\frac{g_w}{\overline{w}}\right\|^2, 1\right\}\right)\right\}.$$

An important final note in this section is that while Theorems 3.3 and 3.5, as stated, require choosing Θ to be some measurable subset of $\mathbb{R}^d$ and that the posterior π concentrates around some point of interest $\theta_0 \in \mathbb{R}^d$, these results can be generalized to a wider class of models and spaces. In particular, Corollary 3.6 demonstrates that if Θ is arbitrary, but the potential functions ℓ_n only depend on θ through some other function $\eta : \Theta \to \mathbb{R}^d$, that the conclusions of Theorems 3.3 and 3.5 still hold.

Corollary 3.6. *Suppose Θ is an arbitrary measurable space, and the potential functions take the form $\ell_n(\eta(\theta))$ for some measurable function $\eta : \Theta \to \mathbb{R}^d$. Then if the assumptions of Theorems 3.3 and 3.5 hold for potentials $(\ell_n)_{n=1}^N$ as functions on $\mathbb{R}^d$ and pushforward prior $\eta\pi_0$ on $\mathbb{R}^d$, the stated lower bounds also hold for* $\min\{\mathrm{KL}(\pi||\pi_w), \mathrm{KL}(\pi_w||\pi)\}$.

4 Lower bound applications

In this section, the general theoretical results from Section 3 are applied to specific algorithms, Bayesian models, and data generating processes to explain previously observed empirical behaviour of coreset construction, as well as to place fundamental limits on the necessary size of coresets. Consider a setting where the data X_n arise as an i.i.d. sequence drawn from some probability distribution ν, $\ell_n(\eta(\theta)) = \log p(X_n|\eta(\theta))$ for $\eta : \Theta \to \mathbb{R}^d$, $\eta_0 = \eta(\theta_0)$, and the following technical criteria hold (where $\mathbb{E}$ denotes expectation under the data generating process):

(A1) $\mathbb{E}\left[\nabla \ell_n(\eta_0)\right] = 0$ and $H = \mathbb{E}\left[-\nabla^2 \ell_n(\eta_0)\right] = \mathbb{E}\left[\nabla \ell_n(\eta_0)\nabla \ell_n(\eta_0)^T\right] \succ 0$.

(A2) $\mathbb{E}\left[\|\nabla \ell_n(\eta_0)\|_2^{2+\delta}\right] < \infty$ for some $\delta > 0$ and $\mathbb{E}\left[\|\nabla^2 \ell_n(\eta_0)\|_F^2\right] < \infty$.

(A3) On a neighbourhood of η_0, $\|\nabla^2 \ell_n(\eta) - \nabla^2 \ell_n(\eta_0)\|_2 \leq R(X_n)\|\eta - \eta_0\|_2$, $\mathbb{E}\left[R(X_n)\right] < \infty$.

(A4) $\eta\pi_0$ is twice differentiable a neighbourhood of η_0, and $\pi(\eta_0) > 0$.

(A5) For all $r \to 0$ such that $r^2 = \omega(\log N/N)$, $-\log \eta\pi(\|\eta - \eta_0\| > r) = \Omega_p(Nr^2)$.

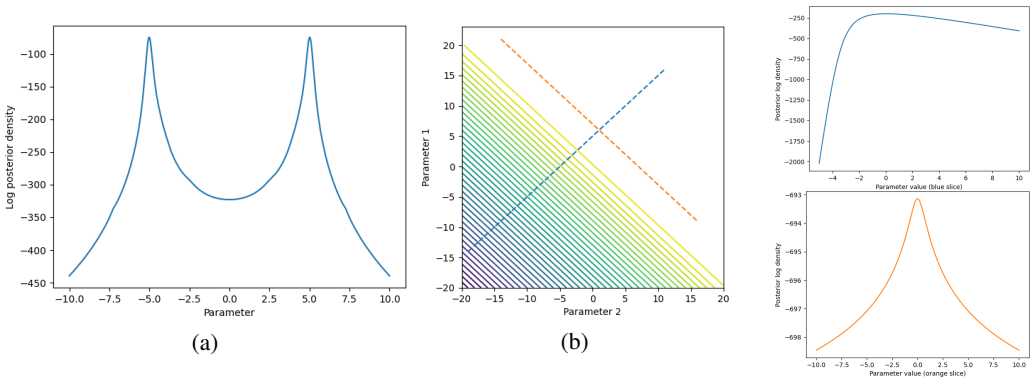
These conditions apply to a wide range of models, e.g., an unidentifiable, multimodal location model posterior with heavy tails on $\Theta = \mathbb{R}$, where the Bayesian model is specified by

$$\theta \sim \mathrm{Cauchy}(0, 1) \qquad (X_n)_{n=1}^N \overset{\text{iid}}{\sim} \mathrm{Cauchy}(\theta^2, 1), \tag{2}$$

and the data are generated from the likelihood with parameter $\theta_0 = 5$, and an unidentifiable logistic regression posterior with heavy tails on $\mathbb{R}^2$, where the Bayesian model is specified by

$$\theta \sim \mathrm{Cauchy}(0, I) \qquad Y_n \overset{\text{ind}}{\sim} \mathrm{Bern}\left(\frac{1}{1 + e^{-X_n^T A\theta}}\right) \qquad A = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}, \tag{3}$$

the covariates are generated via $X_n \overset{\text{iid}}{\sim} \mathrm{Unif}(\{x \in \mathbb{R}^2 : \|x\|_2 \leq 1\})$, and the observations Y_n are generated from the likelihood with parameter $\theta_0 = [1 \quad 6]^T$. See Proposition A.6 in Appendix A for the verification of (A1-5) for these two models. Example posterior log-densities for these models are displayed in Fig. 1.



(a) (b)

Figure 1: Example unnormalized posterior densities given 50 data points for (1a) the Cauchy location model and (1b) the logistic regression model. The orange and blue dashed lines in (1b) indicate one-dimensional slices that are shown in the rightmost panels.

Algorithm 1 Importance-weighted coreset construction

Compute probabilities $(p_n)_{n=1}^N$ (may depend on the data and model)
Draw $I_1, \ldots, I_M \overset{\text{iid}}{\sim} \text{Categorical}(p_1, \ldots, p_N)$
For each n, set $w_n = \frac{1}{Mp_n} \sum_{m=1}^M \mathbb{1}[I_m = n]$.
return $(w_n)_{n=1}^N$

Algorithm 2 Scaled importance-weighted coreset construction

Obtain coreset weights $(w_n)_{n=1}^N$ via Algorithm 1
Compute $\alpha^\star = \arg\min_{\alpha \geq 0} \text{KL}(\pi_{\alpha w} || \pi)$
return $(\alpha^\star w_n)_{n=1}^N$

4.1 Minimum coreset size for importance-weighted coresets

A popular algorithm for coreset construction that has appeared in a wide variety of domains—e.g., Bayesian inference [32, 33, Section 4.1], frequentist inference (e.g., [41–45]), and optimization (see [46] for a recent survey)—involves subsampling of the data followed by an importance-weighting correction. The pseudocode is given in Algorithm 1. Note that $\mathbb{E}[w_n] = 1$, and so $\mathbb{E}[\ell_w] = \ell$; the coreset potential is an unbiased estimate of the exact potential. The advantage of this method is that it is straightforward and computationally efficient. If the sampling probabilities are uniform $p_n = 1/N$, then Algorithm 1 constructs a coreset in $O(M)$ time and $O(M)$ memory. Nonuniform probabilities p_n require $\Omega(N)$ time, as they require a pass over all N data points to compute each p_n [32, 42] followed by sampling the coreset, e.g., via an alias table [47, 48]. However, empirical results produced by this methodology have generally been underwhelming, even with carefully chosen sampling probabilities; see, e.g., Figure 2 of [32].

Corollary 4.1 explains these poor results: Bayesian coresets constructed via Algorithm 1 must satisfy $M \propto N$ in order to maintain a bounded $\underline{\text{KL}}(w)$ in the data-asymptotic limit. In other words, such coresets do not satisfy the desiderata in Section 2. The only restriction is that there exist constants $c, C > 0$ such that for all $N \in \mathbb{N}$, the sampling probabilities $(p_n)_{n=1}^N$ satisfy

$$\text{(A6)} \qquad 0 < c \leq \min_n Np_n \leq \max_n Np_n \leq C < \infty \quad a.s. \tag{4}$$

The lower threshold ensures that the variance of the importance-weighted log-likelihood is not too large, while the upper threshold ensures sufficient diversity in the draws from subsampling. The condition in Eq. (4) is not a major restriction, in the sense that performance should deteriorate even further when it does not hold. The $(p_n)_{n=1}^N$ may otherwise depend arbitrarily on the data and model.

Corollary 4.1. *Given (A1-6), $M \to \infty$, and $M = o(N)$, coresets produced by Algorithm 1 satisfy*

$$\underline{\text{KL}}(w) = \Omega_p\left(\frac{N}{M}\right). \tag{5}$$

The intuition behind Corollary 4.1 is that both the true posterior and the importance-weighted coreset posterior are asymptotically approximately normal with variance $\propto 1/N$ as $N \to \infty$; however, the coreset posterior mean is roughly $\propto M^{-1/2}$ away from the posterior mean, because the subsample is of size M. The KL divergence between two Gaussians is lower-bounded by the inverse variance times the mean difference squared, yielding $\approx N/M$ as in Eq. (5).

Given the intuition that the coreset posterior mean is far from the posterior mean relative to their variances, it is worth asking whether one can apply a small amount of effort to "correct" the importance-weighted coreset by scaling the weights (and hence the variance) down, as shown in Algorithm 2. Unfortunately, Corollary 4.2 demonstrates that even with optimal scaling, $M \propto N$ is still required in order to maintain a bounded KL divergence as $N \to \infty$.

Corollary 4.2. *Given (A1-6), $M \to \infty$, and $M = o(N)$, coresets produced by Algorithm 1 satisfy*

$$\min_{\alpha > 0} \underline{\mathrm{KL}}(\alpha w) = \Omega_p\left(\log \frac{N}{M}\right).$$

Fig. 2 provides empirical confirmation of Corollaries 4.1 and 4.2 on the Cauchy location and logistic regression models in Eqs. (2) and (3). In particular, these figures show that the empirical rates of growth of KL as a function of N closely matches $\Omega_p(\frac{N}{M})$ for importance-weighted coresets, and $\Omega_p(\log \frac{N}{M})$ for the same with post-hoc scaling, for a wide range of coreset sizes $M \in \{\log N, \sqrt{N}, 1/2N\}$. Thus, importance weighted coreset construction methods do not satisfy the desiderata in Section 2 for a wide range of models, and alternate methods should be considered.

4.2 Minimum coreset size for any coreset construction

This section extends the minimum coreset size results from importance-weighted schemes to *any* coreset construction algorithm. In particular, Corollary 4.3 shows that under (A7)—a strengthening of (A3) and Assumption 3.4—and (A8)—which asserts that $\nabla \ell_1(\eta_0), \ldots, \nabla \ell_M(\eta_0)$ are linearly independent a.s. and satisfy a technical moment condition—at least d coreset points are required to keep the KL divergence bounded as $N \to \infty$.

(A7) Assumption 3.4 holds and there exists $\gamma > 0$ such that for all sufficiently large $N \in \mathbb{N}$,

$$\forall \eta \in B, n \in [N], \quad -\nabla^2 \ell_n(\eta) \succeq \gamma H \quad \text{and} \quad L_n^2 < \gamma^{-1} L^2.$$

(A8) For all coreset sizes $M < d$, there exists a $\delta > 0$ such that

$$\mathbb{E}\Big[\big(1^T (G^T G)^{-1} 1\big)^{M+\delta}\Big] < \infty \qquad G = [\nabla \ell_1(\eta_0) \quad \ldots \quad \nabla \ell_M(\eta_0)] \in \mathbb{R}^{d \times M}.$$

Corollary 4.3. *For a fixed coreset size $M < d$, given (A1-5,7,8),*

$$\min_{w \in \mathbb{R}_+^N : \|w\|_0 \leq M} \underline{\mathrm{KL}}(w) = \Omega_p(\log N).$$

5 Upper bounds on approximation error

This section presents upper bounds on the KL divergence of coreset approximations. As in Section 3, the first step is to write all expectations in terms of distributions that do not depend on w. Lemma 5.1 does so without imposing any major assumptions; the result again applies generally in any setting that a Bayesian coreset can be used. For convenience, define

$$\overline{\mathrm{KL}}(w) := \max\{\mathrm{KL}(\pi_w || \pi), \mathrm{KL}(\pi || \pi_w)\}.$$

Lemma 5.1 (Basic KL Upper Bound)**.** *For all coreset weights w,*

$$\overline{\mathrm{KL}}(w) \leq \inf_{\lambda > 0} \frac{1}{\lambda} \log \int \pi \exp\big((1+\lambda)(\bar{\ell}_w - \bar{\ell})\big),$$

where for all $n \in [N]$, $\bar{\ell}_n = \ell_n - \int \pi \ell_n$, $\bar{\ell} = \sum_n \bar{\ell}_n$, and $\bar{\ell}_w = \sum_n w_n \bar{\ell}_n$.

The upper bound in Lemma 5.1 is nonvacuous (i.e., finite) as long as there exists a $\alpha > 1$ such that the α Rényi divergence $D_\alpha(\pi_w || \pi)$ [49, p. 3799] is finite. Note that as in Lemma 3.1, the bound in

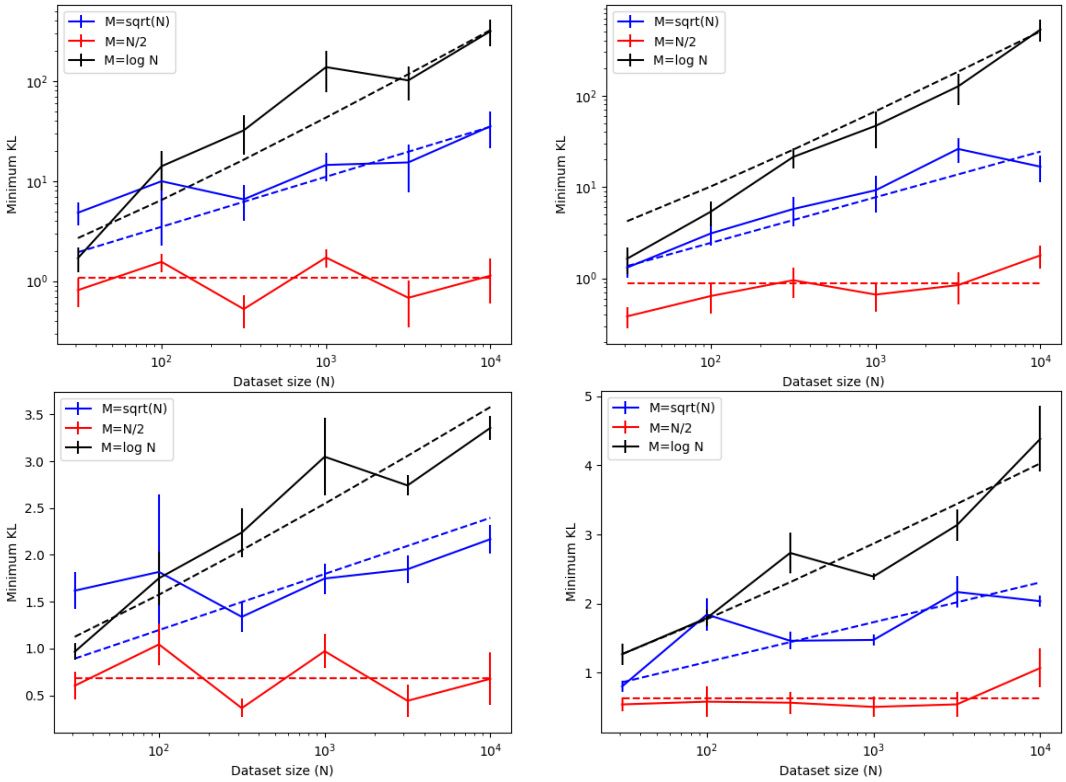


Figure 2: Importance-weighted coreset quality, showing the minimum of the forward and reverse KL divergences on the vertical axis as a function of dataset size N for 3 coreset sizes: $\log N$ (black), $\sqrt{N}$ (blue), and $1/2N$ (red). Dashed lines indicate predictions from the theory in Corollaries 4.1 and 4.2, solid lines indicate the mean over 10 trials, and error bars indicate standard error. The top row shows the quality of basic importance-weighted coresets (note that both horizontal and vertical axes are in log scale), while the bottom row shows the quality with optimal post-hoc scaling (note that only the horizontal axis is in log scale). The left column corresponds to the Cauchy location model, while the right column corresponds to the logistic regression model. Sampling probabilities p_n for both models are set proportional to X_n^2, thresholded to lie between $0.1/N$ and $10/N$.

Lemma 5.1 remains valid if one replaces ℓ_w with $\ell_w - c$ and ℓ with $\ell - c'$ for any constants c, c' that do not depend on θ but may depend on the coreset weights w and data.

More practical bounds necessitate an assumption about the behaviour of the potentials $(\ell_n)_{n=1}^N$. Definition 5.2 below asserts that the multivariate moment generating function of $(\ell_n)_{n=1}^N$ is bounded when the vector is close to 0. This definition is a generalization of the usual definition of subexponentiality for the univariate setting (e.g., [50, Sec. 2.7]). Theorem 5.3 subsequently shows that Definition 5.2 is sufficient to obtain simple bounds on $\overline{\mathrm{KL}}$.

Definition 5.2. *For $A \in \mathbb{R}^{N\times N}$, $A \succeq 0$, and monotone function $h : \mathbb{R}_+ \to \mathbb{R}_+$ such that $\lim_{x\to 0} h(x) = h(0) = 0$, the potentials $(\ell_n)_{n=1}^N$ are (h, A)-*subexponential *if*

$$\forall w \in \mathbb{R}^N : w^T A w \leq 1, \qquad \int \pi \exp\bigl(\bar{\ell}_w\bigr) \leq \exp\bigl(h(w^T A w)\bigr).$$

Theorem 5.3. *If the potentials $(\ell_n)_{n=1}^N$ are (h, A)-subexponential, then*

$$\forall w \in \mathbb{R}_+^N : 4(w-1)^T A (w-1) \leq 1, \qquad \overline{\mathrm{KL}}(w) \leq h(4(w-1)^T A (w-1)).$$

Definition 5.2, the key assumption in Theorem 5.3, is satisfied by a wide range of models when choosing $h(x) = x$ and $A \propto \mathrm{Cov}_\pi\bigl((\ell_n)_{n=1}^N\bigr)$, as demonstrated by Proposition 5.4. Because this case applies widely, let *A-subexponential* be shorthand for (h, A)-subexponentiality with $h(x) = x$.

Proposition 5.4. *If for all w in a ball centered at the origin, $\int \pi \exp(\bar{\ell}_w) < \infty$, then there exists $\beta > 0$ such that the potentials $(\ell_n)_{n=1}^N$ are $\beta\,\mathrm{Cov}_\pi\bigl((\ell_n)_{n=1}^N\bigr)$-subexponential.*

Algorithm 3 Subsample-optimize coreset construction

Compute probabilities $(p_n)_{n=1}^N$ (may depend on the data and model)
Draw $I_1, \dots, I_M \overset{\text{iid}}{\sim} \text{Categorical}(p_1, \dots, p_N)$, and set $\mathcal{I} = \{I_1, \dots, I_M\}$
Compute $w^\star = \arg\min_{w \in \mathbb{R}_+^N} \text{KL}(\pi_w || \pi) \quad \text{s.t.} \quad w_n \neq 0 \text{ only if } n \in \mathcal{I}$.
return $(w_n^\star)_{n=1}^N$

In other words, intuitively, if a coreset construction algorithm produces weights such that $\text{Var}_\pi(\bar{\ell}_w - \bar{\ell})$ is small, then $\overline{\text{KL}}(w)$ is also small. That being said, the generality of Definition 5.2 to allow arbitrary h, A is still helpful in obtaining upper bounds in specific cases; see, e.g., Propositions A.1 and A.2.

6 Upper bound application: subsample-optimize coresets

A strategy to construct Bayesian coresets that has recently emerged in the literature, shown in Algorithm 3, is to first subsample the data to select M data points, and subsequently optimize the weights for those selected data points [36–38]. The subsampling step serves to pick a reasonably flexible basis of log-likelihood functions for coreset approximation, and avoids the slow greedy selection routines from earlier work [33–35]. The optimization step tunes the weights for the selected basis, avoiding the poor approximations of importance-weighting methods. Indeed, Algorithm 3 creates exact coresets $\pi_{w^\star} = \pi$ with high probability in Gaussian location models [36, Prop. 3.1] and finite-dimensional exponential family models [37, Thm. 4.1], and near-exact coresets with high probability in strongly log-concave models [37, Thm. 4.2] and Bayesian linear regression [38, Prop. 3].

Corollary 6.1 generalizes these results substantially, and demonstrates that coresets of size $M = O(\text{polylog}(N))$ produced by the subsample-optimize method in Algorithm 3 maintain a bounded KL divergence as $N \to \infty$. Two key assumptions are subexponentiality of the potentials and a polynomial (in N) growth of $\text{Var}_\pi(\ell(\theta))$; these conditions are not stringent and should hold for a wide range of Bayesian models and i.i.d. data generating processes. The last key assumption in Eq. (6) is that a randomly-chosen potential function ℓ_I, $I \sim \text{Categorical}(p_1, \dots, p_N)$ (with probabilities as in Algorithm 3) is well-aligned with the residual coreset error function. Similar alignment conditions have appeared in past results for more restrictive settings (see, e.g., $J(\delta)$ in [37, Thm. 4.1]).

Corollary 6.1. *Suppose there exist $\beta, \alpha > 0$ and $0 \leq \rho, \epsilon < 1$ such that the potential functions $(\ell_n)_{n=1}^N$ are $\beta \,\text{Cov}_\pi((\ell_n)_{n=1}^N)$-subexponential with probability increasing to 1 as $N \to \infty$, $\text{Var}_\pi(\ell(\theta)) = O_p(N^\alpha)$, $M = (\log N)^{\frac{1}{1-\rho}}$, and*

$$\mathbb{P}\Big(\max\big\{0, \text{Corr}_\pi\big(\ell_{I_M}(\theta), \ell(\theta) - \ell^\star_{M-1}(\theta)\big)\big\}^2 \geq 1 - \epsilon \Big| (\ell_n)_{n=1}^N\Big) = \omega_p(M^{-\rho}) \tag{6}$$

$$\ell^\star_{M-1}(\theta) = \underset{g \in \text{cone}\{\ell_{I_1}, \dots, \ell_{I_{M-1}}\}}{\arg\min} \text{Var}_\pi(\ell(\theta) - g(\theta)) \qquad I_1, \dots, I_M \overset{iid}{\sim} \text{Categorical}(p_1, \dots, p_N).$$

Then Algorithm 3 produces a coreset with $\overline{\text{KL}}(w) = O_p(1)$ as $N \to \infty$.

Fig. 3 confirms that subsample-optimize coreset construction methods applied to the logistic regression and Cauchy location models in Eqs. (2) and (3) (which both violate the conditions of past upper bounds in the literature) are able to provide high-quality posterior approximations for very small coresets—in this case, $M \propto \log N$.

7 Conclusions

This article presented new general lower and upper bounds on the quality of Bayesian coreset approximations, as measured by the KL divergence. These results were used to draw novel conclusions regarding importance-weighted and subsample-optimize coreset methods, which align with simulation experiments on two synthetic models that violate the assumptions of past theoretical results. Avenues for future work include general bounds on the subexponentiality constant β in Proposition 5.4, as well as the alignment probability in Eq. (6), in the setting of Bayesian models with i.i.d. data generating processes. A limitation of this work is that both quantities currently require case-by-case analysis.

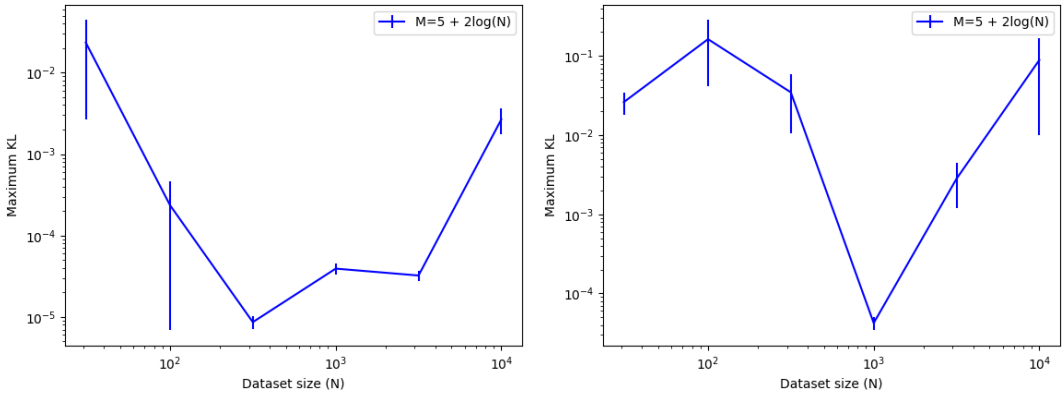


Figure 3: Subsample-optimize coreset quality, showing the maximum of the forward and reverse KL divergences on the vertical axis as a function of dataset size N for coresets of size $5 + 2 \log N$. Solid lines indicate the mean over 70 trials, and error bars indicate standard error. The left panel is for the Cauchy location model, while the right panel is for the logistic regression model. Sampling probabilities are uniform $p_n = 1/N$, and coreset weights were optimized by nonnegative least squares for log-likelihoods discretized via samples from π [34, Eq. 4].

Acknowledgments and Disclosure of Funding

The author gratefully acknowledges the support of an NSERC Discovery Grant (RGPIN-2019-03962).

References

[1] Christian Robert and George Casella. *Monte Carlo Statistical Methods*. Springer, 2nd edition, 2004.

[2] Andrew Gelman, John Carlin, Hal Stern, David Dunson, Aki Vehtari, and Donald Rubin. *Bayesian data analysis*. CRC Press, 3rd edition, 2013.

[3] J. Andrés Christen and Colin Fox. Markov chain Monte Carlo using an approximation. *Journal of Computational and Graphical Statistics*, 14(4):795–810, 2005.

[4] Marco Banterle, Clara Grazian, Anthony Lee, and Christian P. Robert. Accelerating Metropolis-Hastings algorithms by delayed acceptance. *Foundations of Data Science*, 1(2):103–128, 2019.

[5] Richard Payne and Bani Mallick. Bayesian big data classification: a review with complements. *arXiv:1411.5653*, 2014.

[6] Chris Sherlock, Andrew Golightly, and Daniel Henderson. Adaptive, delayed-acceptance MCMC for targets with expensive likelihoods. *Journal of Computational and Graphical Statistics*, 26(2):434–444, 2017.

[7] Arnaud Doucet, Michael Pitt, George Deligiannidis, and Robert Kohn. Efficient implementation of Markov chain Monte Carlo when using an unbiased likelihood estimator. *Biometrika*, 102(2):295–313, 2015.

[8] Dougal Maclaurin and Ryan Adams. Firefly Monte Carlo: exact MCMC with subsets of data. In *Conference on Uncertainty in Artificial Intelligence*, 2014.

[9] Matias Quiroz, Minh-Ngoc Tran, Mattias Villani, Robert Kohn, and Khue-Dung Dang. The block-Poisson estimator for optimally tuned exact subsampling MCMC. *Journal of Computational and Graphical Statistics*, 30(4):877–888, 2021.

[10] Max Welling and Yee Whye Teh. Bayesian learning via stochastic gradient Langevin dynamics. In *International Conference on Machine Learning*, 2011.

[11] Sungjin Ahn, Anoop Korattikara, and Max Welling. Bayesian posterior sampling via stochastic gradient Fisher scoring. In *International Conference on Machine Learning*, 2012.

[12] Anoop Korattikara, Yutian Chen, and Max Welling. Austerity in MCMC land: cutting the Metropolis-Hastings budget. In *International Conference on Machine Learning*, 2014.

[13] Tianqi Chen, Emily Fox, and Carlos Guestrin. Stochastic gradient Hamiltonian Monte Carlo. In *International Conference on Machine Learning*, 2015.

[14] James Johndrow, Natesh Pillai, and Aaron Smith. No free lunch for approximate MCMC. *arXiv:2010.12514*, 2020.

[15] Rémi Bardenet, Arnaud Doucet, and Chris Holmes. On Markov chain Monte Carlo methods for tall data. *Journal of Machine Learning Research*, 18:1–43, 2017.

[16] Tigran Nagapetyan, Andrew Duncan, Leonard Hasenclever, Sebastian Vollmer, Lukasz Szpruch, and Konstantinos Zygalakis. The true cost of stochastic gradient Langevin dynamics. *arXiv:1706.02692*, 2017.

[17] Jack Baker, Paul Fearnhead, Emily Fox, and Christopher Nemeth. Control variates for stochastic gradient MCMC. *Statistics and Computing*, 29:599–615, 2019.

[18] Christopher Nemeth and Paul Fearnhead. Stochastic gradient Markov Chain Monte Carlo. *Journal of the American Statistical Association*, 116(533):433–450, 2021.

[19] Matias Quiroz, Robert Kohn, Mattias Villani, and Minh-Ngoc Tran. Speeding up MCMC by efficient data subsampling. *Journal of the American Statistical Association*, 114(526):831–843, 2019.

[20] Matias Quiroz, Robert Kohn, and Khue-Dung Dang. Subsampling MCMC—an introduction for the survey statistician. *Sankhya: The Indian Journal of Statistics*, 80-A:S33–S69, 2018.

[21] David Blei, Alp Kucukelbir, and Jon McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.

[22] Zhenming Shun and Peter McCullagh. Laplace approximation of high dimensional integrals. *Journal of the Royal Statistical Society: Series B*, 57(4):749–760, 1995.

[23] Peter Hall, Tung Pham, Matt Wand, and Shen S.J. Wang. Asymptotic normality and valid inference for gaussian variational approximation. *The Annals of Statistics*, 39(5):2502–2532, 2011.

[24] Aad van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 2000.

[25] Yixin Wang and David Blei. Frequentist consistency of variational Bayes. *Journal of the American Statistical Association*, 114(527):1147–1161, 2018.

[26] Badr-Eddine Chérief-Abdellatif and Pierre Alquier. Consistency of variational Bayes inference for estimation and model selection in mixtures. *Electronic Journal of Statistics*, 12:2995–3035, 2018.

[27] Yun Yang, Debdeep Pati, and Anirban Bhattacharya. α-variational inference with statistical guarantees. *The Annals of Statistics*, 2018.

[28] Pierre Alquier and James Ridgway. Concentration of tempered posteriors and of their variational approximations. *The Annals of Statistics*, 48(3):1475–1497, 2020.

[29] Zuheng Xu and Trevor Campbell. The computational asymptotics of Gaussian variational inference and the Laplace approximation. *Statistics and Computing*, 32(63), 2022.

[30] Jeffrey Miller. Asymptotic normality, concentration, and coverage of generalized posteriors. *Journal of Machine Learning Research*, 22:1–53, 2021.

[31] Cheng Zhang, Judith Bütepage, Hedvig Kjellström, and Stephan Mandt. Advances in variational inference. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(8):2008–2026, 2019.

[32] Jonathan Huggins, Trevor Campbell, and Tamara Broderick. Coresets for scalable Bayesian logistic regression. In *Advances in Neural Information Processing Systems*, 2016.

[33] Trevor Campbell and Tamara Broderick. Automated scalable Bayesian inference via Hilbert coresets. *Journal of Machine Learning Research*, 20(15):1–38, 2019.

[34] Trevor Campbell and Boyan Beronov. Sparse variational inference: Bayesian coresets from scratch. In *Advances in Neural Information Processing Systems*, 2019.

[35] Trevor Campbell and Tamara Broderick. Bayesian coreset construction via greedy iterative geodesic ascent. In *International Conference on Machine Learning*, 2018.

[36] Naitong Chen, Zuheng Xu, and Trevor Campbell. Bayesian inference via sparse Hamiltonian flows. In *Advances in Neural Information Processing Systems*, 2022.

[37] Cian Naik, Judith Rousseau, and Trevor Campbell. Fast Bayesian coresets via subsampling and quasi-Newton refinement. In *Advances in Neural Information Processing Systems*, 2022.

[38] Martin Jankowiak and Du Phan. Surrogate likelihoods for variational annealed importance sampling. In *International Conference on Machine Learning*, 2022.

[39] Naitong Chen and Trevor Campbell. Coreset Markov chain Monte Carlo. In *International Conference on Artificial Intelligence and Statistics*, 2024.

[40] Thomas Cormen, Charles Leiserson, Ronald Rivest, and Clifford Stein. *Introduction to Algorithms*. The MIT Press, 4th edition, 2022.

[41] Ping Ma, Michael Mahoney, and Bin Yu. A statistical perspective on algorithmic leveraging. *Journal of Machine Learning Research*, 16:861–911, 2015.

[42] HaiYing Wang, Rong Zhu, and Ping Ma. Optimal subsampling for large sample logistic regression. *Journal of the American Statistical Association*, 113(522):829–844, 2018.

[43] HaiYing Wang. More efficient estimation for logistic regression with optimal subsamples. *Journal of Machine Learning Research*, 20:1–59, 2019.

[44] Mingyao Ai, Jun Yu, Huiming Zhang, and HaiYing Wang. Optimal subsampling algorithms for big data regressions. *Statistica Sinica*, 31(2):749–772, 2021.

[45] HaiYing Wang and Yanyuan Ma. Optimal subsampling for quantile regression in big data. *Biometrika*, 108(1):99–112, 2021.

[46] Dan Feldman. Introduction to core-sets: an updated survey. *arXiv:2011.09384*, 2020.

[47] Alastair Walker. New fast method for generating discrete random numbers with arbitrary frequency distributions. *Electronics Letters*, 10(8):127–128, 1974.

[48] Alastair Walker. An efficient method for generating discrete random variables with general distributions. *ACM Transactions on Mathematical Software*, 3(3):253–256, 1977.

[49] Tim van Erven and Peter Harrëmos. Rényi divergence and Kullback-Leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820, 2014.

[50] Roman Vershynin. *High-dimensional probability: an introduction with applications in data science*. Cambridge University Press, 2020.

[51] Igor Vajda. Note on discrimination information and variation. *IEEE Transactions on Information Theory*, 16(6):771–773, 1970.

[52] David Pollard. *A user's guide to probability theory*. Cambridge series in statistical and probabilistic mathematics. Cambridge University Press, 7th edition, 2002.

[53] Robert Keener. *Theoretical statistics: topics for a core course*. Springer, 2010.

[54] Andre Bulinski. Conditional central limit theorem. *Theory of Probability & its Applications*, 61(4):613–631, 2017.

[55] Lorraine Schwartz. On Bayes procedures. *Z. Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 4:10–26, 1965.

A Proofs

Proof of Lemma 3.1. By Vajda's inequality [51],

$$\begin{aligned}\underline{\mathrm{KL}}(w) &\geq \log \frac{1+\mathrm{TV}(\pi,\pi_w)}{1-\mathrm{TV}(\pi,\pi_w)} - \frac{2\,\mathrm{TV}(\pi,\pi_w)}{1+\mathrm{TV}(\pi,\pi_w)} \\ &\geq -\log\left(1-\mathrm{TV}(\pi,\pi_w)\right) - \mathrm{TV}(\pi,\pi_w) \\ &\geq 0.\end{aligned}$$

The bound is monotone increasing in $\mathrm{TV}(\pi,\pi_w)$; therefore because the squared Hellinger distance satisfies the inequality [52, p. 61],

$$H^2(\pi,\pi_w) = \frac{1}{2}\int \left(\sqrt{\pi} - \sqrt{\pi_w}\right)^2 \leq \frac{1}{2}\int |\pi - \pi_w| = \mathrm{TV}(\pi,\pi_w),$$

we have that

$$\underline{\mathrm{KL}}(w) \geq -\log\left(1 - H^2(\pi,\pi_w)\right) - H^2(\pi,\pi_w).$$

We substitute the value of the squared Hellinger distance to find that

$$\underline{\mathrm{KL}}(w) \geq -\log\left(\int \sqrt{\pi\pi_w}\right) + \int \sqrt{\pi\pi_w} - 1 \geq 0.$$

Note that $\int \sqrt{\pi\pi_w} \leq 1$, so

$$\underline{\mathrm{KL}}(w) \geq -\log\left(\min\{1, \int \sqrt{\pi\pi_w}\}\right) + \min\{1, \int \sqrt{\pi\pi_w}\} - 1 \geq 0.$$

The bound is monotone decreasing in $\int \sqrt{\pi\pi_w}$, so we require an upper bound on $\int \sqrt{\pi\pi_w}$. To obtain the required bound, we split the integral into two parts—one on the set B, and the other on B^c—and then use the Cauchy-Schwarz inequality to bound the part on B^c. Note that by definition π and π_w are mutually dominating, so the density ratio π_w/π is well-defined and measurable.

$$\begin{aligned}\int \sqrt{\pi\pi_w} &= \int_B \sqrt{\pi\pi_w} + \int_{B^c} \sqrt{\pi\pi_w} \\ &= \int_B \sqrt{\pi\pi_w} + \int \pi \sqrt{\frac{\pi_w}{\pi}}\mathbb{1}_{B^c} \\ &\leq \int_B \sqrt{\pi\pi_w} + \sqrt{\pi(B^c)} \\ &= \frac{\int_B \pi_0 \exp\frac{1}{2}(\ell+\ell_w)}{\sqrt{\int \pi_0 \exp(\ell)\int \pi_0 \exp(\ell_w)}} + \sqrt{\pi(B^c)}.\end{aligned}$$

The result follows. □

Proof of Lemma 5.1. We first consider the forward KL divergence. By definition,

$$\begin{aligned}\mathrm{KL}(\pi||\pi_w) &= \int \pi(\ell-\ell_w) + \log\frac{\int \pi_0\exp(\ell_w)}{\int \pi_0\exp(\ell)} \\ &= \int \pi(\ell-\ell_w) + \log\int \pi\exp(\ell_w-\ell).\end{aligned}$$

Since the KL is positive, for $\lambda > 0$,

$$\begin{aligned}\mathrm{KL}(\pi||\pi_w) &\leq \frac{1+\lambda}{\lambda}\int \pi(\ell-\ell_w) + \frac{1+\lambda}{\lambda}\log\int \pi\exp(\ell_w-\ell) \\ &\leq \frac{1+\lambda}{\lambda}\int \pi(\ell-\ell_w) + \frac{1}{\lambda}\log\int \pi\exp((1+\lambda)(\ell_w-\ell)) \\ &= \frac{1}{\lambda}\log\int \pi\exp((1+\lambda)(\bar{\ell}_w-\bar{\ell})),\end{aligned}$$

by Jensen's inequality. Next we consider the reverse KL divergence. For any $\lambda \neq 0$,

$$\begin{aligned}\mathrm{KL}(\pi_w||\pi) &= \int \pi_w(\ell_w-\ell) + \log\frac{\int \pi_0\exp(\ell)}{\int \pi_0\exp(\ell_w)} \\ &= \frac{1}{\lambda}\int \pi_w\lambda(\ell_w-\ell) - \log\int \pi\exp(\ell_w-\ell).\end{aligned}$$

By Jensen's inequality, for $\lambda > 0$,

$$\begin{aligned}
\mathrm{KL}(\pi_w||\pi) &\le \frac{1}{\lambda}\log\int \pi_w \exp(\lambda(\ell_w - \ell)) - \log\int \pi\exp(\ell_w - \ell)\\
&= \frac{1}{\lambda}\log\frac{\int \pi\exp((1+\lambda)(\ell_w-\ell))}{\int \pi\exp(\ell_w-\ell)} - \log\int\pi\exp(\ell_w-\ell)\\
&= \frac{1}{\lambda}\log\int\pi\exp((1+\lambda)(\ell_w-\ell)) - \frac{1+\lambda}{\lambda}\log\int\pi\exp(\ell_w-\ell)\\
&\le \frac{1+\lambda}{\lambda}\int\pi(\ell-\ell_w) + \frac{1}{\lambda}\log\int\pi\exp((1+\lambda)(\ell_w-\ell))\\
&= \frac{1}{\lambda}\log\int\pi\exp((1+\lambda)(\bar{\ell}_w-\bar{\ell})).
\end{aligned}$$

This is the same bound as in the forward KL divergence case. Since the bound applies for all $\lambda > 0$, we can take the infimum. □

Proof of Theorem 3.3. By replacing the integrals over the whole space Θ in the denominator of $J_B(w)$ in Lemma 3.1 with integrals over the subset B,

$$\begin{aligned}
\underline{\mathrm{KL}}(w) &\ge -\log\min(1, J_B(w)) + \min(1, J_B(w)) - 1\\
&\ge O_p(1) - \log J_B(w)\\
&\ge O_p(1) + \min\Big\{G_B(w), -\log\sqrt{\pi(B^c)}\Big\}\\
G_B(w) &= -\log\int_B \pi_0\exp((1/2)(\ell+\ell_w)) + \frac{1}{2}\log\int_B\pi_0\exp(\ell) + \frac{1}{2}\log\int_B\pi_0\exp(\ell_w).
\end{aligned}$$

So to obtain the stated lower bound on the KL divergence, we require an upper bound on $\log\int_B \pi_0\exp((1/2)(\ell+\ell_w))$, and lower bounds on $\log\int_B\pi_0\exp(\ell)$ and $\log\int_B\pi_0\exp(\ell_w)$. By Taylor's theorem, Assumption 3.2, and the assumption on $\nabla^2\ell_w(\theta)$, for all $\theta\in B$,

$$\begin{aligned}
\left|\ell(\theta) - \ell(\theta_0) - g^T(\theta-\theta_0) + \frac{N}{2}(\theta-\theta_0)^T H(\theta-\theta_0)\right| &\le \frac{No_p(1)}{2}(\theta-\theta_0)^T H(\theta-\theta_0)\\
\left|\ell_w(\theta) - \ell_w(\theta_0) - g_w^T(\theta-\theta_0) + \frac{\overline{w}}{2}(\theta-\theta_0)^T H(\theta-\theta_0)\right| &\le \frac{\overline{w}o_p(1)}{2}(\theta-\theta_0)^T H(\theta-\theta_0).
\end{aligned} \tag{7}$$

We shift the exponential arguments in $G_B(w)$ by $(1/2)(\ell(\theta_0)+\ell_w(\theta_0))$, note that π_0 is continuous and positive around θ_0, and and apply the Taylor expansions in Eq. (7) to obtain an upper bound on the first term:

$$\log\int_B \pi_0 e^{\frac{1}{2}(\ell-\ell(\theta_0)+\ell_w-\ell_w(\theta_0))} \le O_p(1) + \log\int_B e^{\frac{1}{2}((g+g_w)^T(\theta-\theta_0) - \frac{(\sim 1)(N+\overline{w})}{4}(\theta-\theta_0)^T H(\theta-\theta_0)},$$

where (~ 1) denotes a quantity that converges in probability to 1 as $N\to\infty$. We can transform variables to $x = C^T(\theta-\theta_0)$, where $H = CC^T$ is the Cholesky factorization of H, and subsequently complete the square:

$$\log\int_B \pi_0 e^{\frac{1}{2}(\cdots)} \le O_p(1) + \frac{(\sim 1)\|C^{-1}(g+g_w)\|^2}{4(N+\overline{w})} + \log\int_{\|x\|^2\le r^2} e^{-\frac{(\sim 1)(N+\overline{w})}{4}\left\|x - \frac{(\sim 1)C^{-1}(g+g_w)}{(N+\overline{w})}\right\|^2}. \tag{8}$$

We can obtain lower bounds on the other two terms using a similar technique:

$$\log\int_B \pi_0 e^{\ell-\ell(\theta_0)} \ge O_p(1) + \frac{(\sim 1)\|C^{-1}g\|^2}{2N} + \log\int_{\|x\|^2\le r^2} e^{-\frac{(\sim 1)N}{2}\left\|x - \frac{(\sim 1)C^{-1}g}{N}\right\|^2} \tag{9}$$

$$\log\int_B \pi_0 e^{\ell_w-\ell_w(\theta_0)} \ge O_p(1) + \frac{(\sim 1)\|C^{-1}g_w\|^2}{2\overline{w}} + \log\int_{\|x\|^2\le r^2} e^{-\frac{(\sim 1)\overline{w}}{2}\left\|x - \frac{(\sim 1)C^{-1}g_w}{\overline{w}}\right\|^2}. \tag{10}$$

It remains to analyze the three $\log\int\ldots$ terms. We bound the integral term in Eq. (8) with the integral over the whole space:

$$\log\int_{\|x\|^2\le r^2} e^{-\frac{(\sim 1)(N+\overline{w})}{4}\|\ldots\|^2} \le O_p(1) - \frac{d}{2}\log(N+\overline{w}).$$

For the integral term in Eq. (9), note that since $Nr^2 = \omega(1)$ and $\|C^{-1}g/N\| = O_p(N^{-1/2})$, we have

$$\log\int_{\|x\|^2\le r^2} e^{-\frac{(\sim 1)N}{2}\|\ldots\|^2}$$

$$
\begin{aligned}
&= \log\left(\int e^{-\frac{(\sim 1)N}{2}(\dots)} - \int_{\|x\|^2 > r^2} e^{-\frac{(\sim 1)N}{2}\|x - \frac{(\sim 1)C^{-1}g}{N}\|^2}\right) \\
&\geq \log\left(\left(\frac{(\sim 1)2\pi}{N}\right)^{d/2} - e^{-\frac{(\sim 1)N}{4}\min_{\|x\|\geq r}\left\|x - \frac{(\sim 1)C^{-1}g}{N}\right\|^2} \int e^{-\frac{(\sim 1)N}{4}\|x - \frac{(\sim 1)C^{-1}g}{N}\|^2}\right) \\
&= \log\left(\left(\frac{(\sim 1)2\pi}{N}\right)^{d/2} - e^{-\frac{\Omega_p(Nr^2)}{4}}\left(\frac{(\sim 1)4\pi}{N}\right)^{d/2}\right) \\
&= -\frac{d}{2}\log(N) + O_p(1).
\end{aligned}
$$

For the integral term in Eq. (10), we consider two cases: one where $\overline{w}$ is large, and one where it is small. First assume $\overline{w}r^2 > 8d\log 2$; then by a similar technique as used in the first lower bound, since $\|C^{-1}g_w/\overline{w}\| = o_p(r)$,

$$
\begin{aligned}
&\log \int_{\|x\|^2 \leq r^2} e^{-\frac{(\sim 1)\overline{w}}{2}\|\dots\|^2} \\
&\geq \log\left(\left(\frac{(\sim 1)2\pi}{\overline{w}}\right)^{d/2} - e^{-\frac{(\sim 1)\overline{w}}{4}\min_{\|x\|\geq r}\left\|x - \frac{(\sim 1)C^{-1}g_w}{\overline{w}}\right\|^2} \int e^{-\frac{(\sim 1)\overline{w}}{4}\|x - \frac{(\sim 1)C^{-1}g_w}{\overline{w}}\|^2}\right) \\
&\geq \log\left(\left(\frac{(\sim 1)2\pi}{\overline{w}}\right)^{d/2} - e^{-2d\log 2(\sim 1)}\left(\frac{(\sim 1)4\pi}{\overline{w}}\right)^{d/2}\right) \\
&\geq -\frac{d}{2}\log\overline{w} + O_p(1).
\end{aligned}
$$

When $\overline{w}r^2 \leq 8d\log 2$, we transform variables $y = x/r$ to find that since $\|C^{-1}g_w/\overline{w}\| = o_p(r)$,

$$
\begin{aligned}
\log \int_{\|x\|^2 \leq r^2} e^{-\frac{(\sim 1)\overline{w}}{2}\|\dots\|^2} &= \frac{d}{2}\log r^2 + \log \int_{\|y\|^2 \leq 1} e^{-\frac{(\sim 1)\overline{w}r^2}{2}\left\|y - \frac{(\sim 1)C^{-1}g_w}{r\overline{w}}\right\|^2} \\
&\geq \frac{d}{2}\log r^2 + \log e^{-\frac{8d\log 2(\sim 1)}{2}\left(2 + 2\left\|\frac{(\sim 1)C^{-1}g_w}{r\overline{w}}\right\|^2\right)}\left(\int_{\|y\|^2\leq 1} 1\right) \\
&= \frac{d}{2}\log r^2 + O_p(1).
\end{aligned}
$$

Therefore regardless of the value of $\overline{w}$,

$$
\log \int_{\|x\|^2 \leq r^2} e^{-\frac{(\sim 1)\overline{w}}{2}\|\dots\|^2} \geq -\frac{d}{2}\log\bigl(\max\{\overline{w}, 1/r^2\}\bigr) + O_p(1).
$$

So therefore combining all previous results,

$$
\begin{aligned}
G_B(w) &\geq O_p(1) + \frac{(\sim 1)}{4}\left(\frac{\|C^{-1}g\|^2}{N} + \frac{\|C^{-1}g_w\|^2}{\overline{w}} - \frac{\|C^{-1}(g+g_w)\|^2}{N+\overline{w}}\right) + \frac{d}{4}\log\frac{(N+\overline{w})^2}{N\max\{\overline{w}, 1/r^2\}} \\
&= O_p(1) + \frac{(\sim 1)}{4}\left(\frac{\overline{w}\|C^{-1}g\|^2}{N(N+\overline{w})} + \frac{N\|C^{-1}g_w\|^2}{\overline{w}(N+\overline{w})} - \frac{2g^TH^{-1}g_w}{N+\overline{w}}\right) + \frac{d}{4}\log\frac{(N+\overline{w})^2}{N\max\{\overline{w}, 1/r^2\}} \\
&= O_p(1) + \frac{(\sim 1)}{4}\left(\frac{N\overline{w}}{N+\overline{w}}\left\|\frac{C^{-1}g}{N} - \frac{C^{-1}g_w}{\overline{w}}\right\|^2\right) + \frac{d}{4}\log\frac{(N+\overline{w})^2}{N\max\{\overline{w}, 1/r^2\}} \\
&= O_p(1) + \Omega_p(1)\left(\frac{N\overline{w}}{N+\overline{w}}\left\|\frac{g}{N} - \frac{g_w}{\overline{w}}\right\|^2 + d\log\frac{(N+\overline{w})^2}{N\max\{\overline{w}, 1/r^2\}}\right).
\end{aligned}
$$

We now consider the minimum over $\alpha \geq 0$. Since neither $O_p(1)$ or $\Omega_p(1)$ above depends on $\overline{w}$, we have that

$$
\min_{\alpha\geq 0}\underline{\mathrm{KL}}(\alpha w) \geq O_p(1) + \Omega_p(1)\min\left\{-\log\pi(B^c), \left(\min_{\alpha\geq 0}\frac{N\alpha\overline{w}}{N+\alpha\overline{w}}\left\|\frac{g}{N} - \frac{g_w}{\overline{w}}\right\|^2 + d\log\frac{(N+\alpha\overline{w})^2}{N\max\{\alpha\overline{w}, 1/r^2\}}\right)\right\}.
$$

On the $1/r^2$ branch of the objective function, the derivative in α is always positive, and hence the minimum occurs at $\alpha = 0$, and so

$$
\min_{\alpha\geq 0}(\dots) \geq d\log(Nr^2).
$$

On the αw branch of the objective function,

$$
\min_{\alpha\geq 0}\frac{N\alpha\overline{w}}{N+\alpha\overline{w}}\left\|\frac{g}{N} - \frac{g_w}{\overline{w}}\right\|^2 + d\log\frac{(N+\alpha\overline{w})^2}{N\alpha\overline{w}} \geq \min_{\alpha\geq 0}\frac{N\alpha\overline{w}}{N+\alpha\overline{w}}\left\|\frac{g}{N} - \frac{g_w}{\overline{w}}\right\|^2 + d\log\frac{(N+\alpha\overline{w})}{N\alpha\overline{w}} + d\log N.
$$

For $a, b > 0$ and $x \geq 0$, the function $ax - b\log x$ is convex in x with minimum at $x^\star = b/a$, and so

$$\min_{\alpha \geq 0}(\dots) \geq d \log\left(N\left\|\frac{g}{N} - \frac{g_w}{\overline{w}}\right\|^2\right).$$

By assumption, $\|\frac{g}{N}\| = o_p(r)$ and $\|\frac{g_w}{\overline{w}}\| = o_p(r)$, and hence the αw branch has the asymptotic minimum:

$$\min_{\alpha \geq 0} \underline{\mathrm{KL}}(\alpha w) \geq O_p(1) + \Omega_p(1) \min\left\{-\log \pi(B^c), d\log\left(N\left\|\frac{g}{N} - \frac{g_w}{\overline{w}}\right\|^2\right)\right\}.$$

□

Proof of Theorem 3.5. By Lemma 3.1,

$$\begin{aligned}
\underline{\mathrm{KL}}(w) &\geq -\log \min(1, J_B(w)) + \min(1, J_B(w)) - 1 \\
&\geq O_p(1) + \min\left\{G_B(w), -\log\sqrt{\pi(B^c)}\right\} \\
G_B(w) &= -\log \int_B \pi_0 \exp((1/2)(\ell + \ell_w)) + \frac{1}{2}\log \int \pi_0 \exp(\ell) + \frac{1}{2}\log \int \pi_0 \exp(\ell_w).
\end{aligned}$$

Note that G_B in this proof is subtly different from the G_B used in the proof of Theorem 3.3; the latter two integrals are over the whole space (directly from Lemma 3.1), rather than B. We shift the exponential arguments in $G_B(w)$ by $(1/2)(\ell(\theta_0) + \ell_w(\theta_0))$. We first provide lower bounds on two of the integral terms via Assumption 3.4:

$$\log \int \pi_0 e^{\ell - \ell(\theta_0)} \geq O_p(1) + \log \int e^{(g+g_0)^T(\theta-\theta_0) - \frac{(\sim 1)(N+1)L'^2}{2}\|\theta-\theta_0\|^2},$$

where (~ 1) denotes a quantity that converges in probability to 1, $g_0 = \nabla \log \pi_0(\theta_0)$, and $L'^2 = \frac{NL^2 + L_0^2}{N+1}$. Transforming variables via $x = L'(\theta - \theta_0)$,

$$\begin{aligned}
\log \int \pi_0 e^{\ell - \ell(\theta_0)} &\geq O_p(1) + \log \int e^{(g+g_0)^T x/L' - \frac{(\sim 1)(N+1)}{2}\|x\|^2} \\
&= O_p(1) + \log \int e^{-\frac{(\sim 1)(N+1)}{2}\left\|x - \frac{g+g_0}{(N+1)L'}\right\|^2 + \frac{(\sim 1)(N+1)}{2}\|\frac{g+g_0}{(N+1)L'}\|^2} \\
&= O_p(1) + \frac{(\sim 1)(N+1)}{2L'^2}\left\|\frac{g+g_0}{N+1}\right\|^2 - \frac{d}{2}\log(N+1) \\
&\geq O_p(1) + \frac{(\sim 1)(N+1)}{2\max\{L^2, L_0^2\}}\left\|\frac{g+g_0}{N+1}\right\|^2 - \frac{d}{2}\log(N+1).
\end{aligned}$$

Let $L_w^2 = \frac{1}{\overline{w}}\sum_n w_n L_n^2$. Using the same technique, with $L_w'^2 = \frac{\overline{w}L_w^2 + L_0^2}{\overline{w}+1}$ and $x = L_w'(\theta - \theta_0)$,

$$\begin{aligned}
\log \int \pi_0 e^{\ell_w - \ell_w(\theta_0)} &\geq \log \int e^{(g_w+g_0)^T(\theta-\theta_0) - \frac{(\sim 1)(\overline{w}+1)}{2}L_w'^2\|\theta-\theta_0\|^2} \\
&\geq O_p(1) + \frac{\overline{w}+1}{2L_w'^2}\left\|\frac{g_w+g_0}{\overline{w}+1}\right\|^2 + \log \int e^{-\frac{(\overline{w}+1)}{2}\left\|x - \frac{g_w+g_0}{(\overline{w}+1)L_w'}\right\|^2} \\
&\geq O_p(1) + \frac{\overline{w}+1}{2L_w'^2}\left\|\frac{g_w+g_0}{\overline{w}+1}\right\|^2 + \log \int_{\|x - \frac{g_w+g_0}{(\overline{w}+1)L_w'}\| \leq (\overline{w}+1)^{-1/3}} e^{-\frac{\overline{w}+1}{2}\left\|x - \frac{g_w+g_0}{(\overline{w}+1)L_w'}\right\|^2} \\
&= O_p(1) + \frac{\overline{w}+1}{2L_w'^2}\left\|\frac{g_w+g_0}{\overline{w}+1}\right\|^2 - \frac{d}{2}\log(\overline{w}+1) \\
&\geq O_p(1) + \frac{\overline{w}+1}{2\max\{\beta L^2, L_0^2\}}\left\|\frac{g_w+g_0}{\overline{w}+1}\right\|^2 - \frac{d}{2}\log(\overline{w}+1).
\end{aligned}$$

For the upper bound on the first term, we use a local quadratic expansion around θ_0, where $H_0 = -\nabla^2 \log \pi_0(\theta_0)$,

$$\log \int_B \pi_0 e^{\frac{1}{2}(\ell - \ell(\theta_0) + \ell_w - \ell_w(\theta_0))} \leq O_p(1) + \log \int_B e^{\frac{1}{2}((g+g_w+2g_0)^T(\theta-\theta_0) - \frac{(\sim 1)(N+\overline{w}+2)}{4}(\theta-\theta_0)^T\left(\frac{(N+\alpha\overline{w})H + 2H_0}{N+\overline{w}+2}\right)(\theta-\theta_0)}.$$

Because $H \succ 0$, we have $(N + \alpha\overline{w})H + 2H_0 \succ 0$ eventually; we can transform variables to $x = C^T(\theta - \theta_0)$, where $\frac{(N+\alpha\overline{w})H + 2H_0}{N+\overline{w}+2} = CC^T$ is the Cholesky factorization, and subsequently complete the square. Note that

$$\sqrt{\min\{\min(\alpha, 1)\lambda_{\min}H, \lambda_{\min}H_0\}} \leq \lambda_{\min}C \leq \lambda_{\max}C \leq \sqrt{\max\{\max(\alpha, 1)\lambda_{\max}H, \lambda_{\max}H_0\}}$$

so

$$\log|C| = O_p(1) \qquad \lambda_{\min} C^{-1} H C^{-T} \geq \frac{\lambda_{\min} H}{\max\{\max(\alpha,1)\lambda_{\max} H, \lambda_{\max} H_0\}} = \eta > 0,$$

and therefore

$$\begin{aligned}
&\log \int_B \pi_0 e^{\frac{1}{2}(\dots)} \\
&\leq O_p(1) + \frac{(\sim 1)(N+\overline{w}+2)}{4} \left\| \frac{C^{-1}(g+g_w+2g_0)}{N+\overline{w}+2} \right\|^2 + \log \int_{\|x\|^2 \leq r^2 \eta^{-1}} e^{-\frac{(\sim 1)(N+\overline{w}+2)}{4} \left\| x - \frac{(\sim 1) C^{-1}(g+g_w+2g_0)}{N+\overline{w}+2} \right\|^2}.
\end{aligned} \tag{11}$$

Suppose first that $\overline{w}+1 \leq N/(4\|C^{-1}\|^2 \max\{\beta L^2, L_0^2\})$. In this case we bound the integral in Eq. (11) by integrating over the whole space:

$$\log \int_B \pi_0 e^{\frac{1}{2}(\dots)} \leq O_p(1) + \frac{(\sim 1)\|C^{-1}\|^2 (N+\overline{w}+2)}{4} \left\| \frac{g+g_w+2g_0}{N+\overline{w}+2} \right\|^2 - \frac{d}{2}\log(N+\overline{w}+2).$$

Combining this with the previous results yields

$$\begin{aligned}
&G_B(w) \geq O_p(1) \\
&- \frac{(\sim 1)(N+\overline{w}+2)}{4} \|C^{-1}\|^2 \left\| \frac{g+g_w+2g_0}{N+\overline{w}+2} \right\|^2 \\
&+ \frac{d}{4}\log\frac{(N+\overline{w}+2)^2}{(N+1)(\overline{w}+1)} + \frac{\overline{w}+1}{4\max\{\beta L^2, L_0^2\}} \left\| \frac{g_w+g_0}{\overline{w}+1} \right\|^2 + \frac{(N+1)}{4\max\{L^2, L_0^2\}} \left\| \frac{g+g_0}{N+1} \right\|^2 \\
&\geq O_p(1) + \frac{d}{4}\log\frac{(N+\overline{w}+2)^2}{(N+1)(\overline{w}+1)} + \frac{\overline{w}+1}{4} \left\| \frac{g_w+g_0}{\overline{w}+1} \right\|^2 \left(\frac{1}{\max\{\beta L^2, L_0^2\}} - \frac{2\|C^{-1}\|^2(\overline{w}+1)}{N+\overline{w}+2} \right) \\
&\geq O_p(1) + \frac{d}{4}\log\frac{(N+\overline{w}+2)^2}{(N+1)(\overline{w}+1)} + \frac{\overline{w}+1}{8\max\{\beta L^2, L_0^2\}} \left\| \frac{g_w+g_0}{\overline{w}+1} \right\|^2.
\end{aligned}$$

Bounding the last term below by 0 and minimizing over w such that $\overline{w} \leq \sqrt{N}$ yields

$$G_B(w) \geq O_p(1) + \frac{d}{4}\log\sqrt{N} = O_p(1) + \frac{d}{8}\log N.$$

Bounding $(N+\overline{w}+2)/(N+1) \geq 1$ and minimizing over w such that $\overline{w} \geq \sqrt{N}$ yields

$$\begin{aligned}
G_B(w) &\geq O_p(1) + \frac{d}{4}\log N - \frac{d}{4}\log(\overline{w}+1) + \frac{\overline{w}+1}{8\max\{\beta L^2, L_0^2\}} \left\| \frac{g_w+g_0}{\overline{w}+1} \right\|^2 \\
&\geq O_p(1) + \frac{d}{4}\log N \left\| \frac{g_w+g_0}{\overline{w}+1} \right\|^2 \\
&= O_p(1) + \frac{d}{4}\log N \left\| \frac{g_w}{\overline{w}} \right\|^2,
\end{aligned}$$

where the second line follows because for $a, b > 0$ and $x \geq 0$, the function $ax - b\log x$ is convex in x with minimum at $x^\star = b/a$. Therefore for $\overline{w}+1 \leq N/(\dots)$,

$$\underline{\mathrm{KL}}(w) \geq O_p(1) + \Omega_p(1) d \log\left(N \min\left\{ \left\| \frac{g_w}{\overline{w}} \right\|^2, 1 \right\} \right).$$

Next suppose $\overline{w}+1 \geq N/(4\|C^{-1}\|^2 \max\{\beta L^2, L_0^2\})$. A second upper bound on Eq. (11) can be obtained by taking the maximum of the integrand over the integration region $\|x\|^2 \leq r^2$. Note that since $\|g_w/\overline{w}\| = \omega_p(r)$, $\overline{w} = \Omega_p(N)$, $g/N = O_p(N^{-1/2})$, and $Nr^2 = \omega_p(1)$, we have that $\|(g+g_w+2g_0)/(N+\overline{w}+2)\| = \omega_p(r)$, and so

$$\begin{aligned}
&\log \int_B \pi_0 e^{\frac{1}{2}(\dots)} \\
&\leq O_p(1) + \frac{(\sim 1)(N+\overline{w}+2)}{4} \left\| \frac{C^{-1}(g+g_w+2g_0)}{N+\overline{w}+2} \right\|^2 - \frac{(\sim 1)(N+\overline{w}+2)}{4} \left(\left\| \frac{C^{-1}(g+g_w+2g_0)}{N+\overline{w}+2} \right\| - r \right)^2 + \frac{d}{2}\log r^2 \\
&= O_p(1) - \frac{(\sim 1)(N+\overline{w}+2)}{4} r^2 + \frac{(\sim 1)(N+\overline{w}+2) r}{2} \left\| \frac{C^{-1}(g+g_w+2g_0)}{N+\overline{w}+2} \right\| + \frac{d}{2}\log r^2.
\end{aligned}$$

So therefore combining this result with the previous bounds and minimizing over $\overline{w}$ yields

$$G_B(w) \geq O_p(1) + \frac{(\sim 1)(N+\overline{w}+2)}{4} r^2 - \frac{(\sim 1)(N+\overline{w}+2) r}{2} \left\| \frac{C^{-1}(g+g_w+2g_0)}{N+\overline{w}+2} \right\|$$

$$
\begin{aligned}
&- \frac{d}{4}\log((N+1)(\overline{w}+1)r^4) + \frac{\overline{w}+1}{4\max\{\beta L^2, L_0^2\}}\left\|\frac{g_w+g_0}{\overline{w}+1}\right\|^2 + \frac{(N+1)}{4\max\{L^2, L_0^2\}}\left\|\frac{g+g_0}{N+1}\right\|^2 \\
&\geq O_p(1) - \frac{d}{4}\log(Nr^2) + \frac{(\sim 1)N}{4}\Big(\Big\|\frac{g}{N}\Big\| - r\Big)^2 - \frac{d}{4}\log(\overline{w}r^2) + \frac{(\sim 1)\overline{w}}{4}\Big(\Big\|\frac{g_w}{\overline{w}}\Big\| - r\Big)^2 \\
&\geq O_p(1) - \frac{d}{4}\log(Nr^2) + \frac{(\sim 1)}{4}Nr^2 - \frac{d}{4}\log(r^2) + \frac{d}{4}\log\Big\|\frac{g_w}{\overline{w}}\Big\|^2 \\
&\geq O_p(1) + \frac{d}{4}\log N\Big\|\frac{g_w}{\overline{w}}\Big\|^2.
\end{aligned}
$$

Combining with the earlier bound and noting that $N\min\{\|g_w/\overline{w}\|, 1\} = \omega_p(1)$ yields the final result. □

Proof of Corollary 3.6. The proof follows directly from Theorems 3.3 and 3.5 by the data processing inequality applied to $\underline{\mathrm{KL}}(w)$. □

Proof of Theorem 5.3. By Lemma 5.1,

$$
\begin{aligned}
\overline{\mathrm{KL}}(w) &\leq \inf_{\lambda>0}\frac{1}{\lambda}\log\int \pi\exp\big((1+\lambda)(\bar{\ell}_w - \bar{\ell})\big) \\
&= \inf_{\lambda>0}\frac{1}{\lambda}\log\int \pi\exp\big(\bar{\ell}_{(1+\lambda)(w-1)}\big).
\end{aligned}
$$

Since $(\ell_n)_{n=1}^N$ are (f, A)-subexponential, if

$$
(1+\lambda)^2(w-1)^T A(w-1) \leq 1,
$$

then

$$
\int \pi\exp\big(\bar{\ell}_{(1+\lambda)(w-1)}\big) \leq \exp\Big(f((1+\lambda)^2(w-1)^T A(w-1))\Big).
$$

By assumption, the condition holds when $\lambda = 1$; the result follows. □

Proof of Proposition 5.4. Let $C(w) = \log\int \pi\exp(\bar{\ell}_w)$. By the finiteness condition, [53, Theorem 2.4] asserts that $C(w)$ is continuous, and has derivatives of all orders that can be obtained by passing differentiation through the integral within the set $\|w\|_2 < \alpha$. Let $U = \mathrm{Cov}_\pi((\ell_n)_{n=1}^N)$, and $\mathcal{S} = \mathrm{span}\{w \in \mathbb{R}^N : w^T(\bar{\ell}_n)_{n=1}^N = 0\ \pi\text{-a.s.}\}$. Note that $\mathcal{S} = \ker U$: since $w^T U w = \mathrm{Var}_\pi(w^T(\ell_n)_{n=1}^N)$, $w^T U w = 0$ if and only if $w^T(\bar{\ell}_n)_{n=1}^N = 0\ \pi$-a.s.; and since U is symmetric positive semidefinite, $w^T U w = 0$ if and only if $w \in \ker U$. Therefore $C(w)$ is continuous, has derivatives of all orders, and derivatives can be passed through the integral within the set $\{w \in \mathbb{R}^N : w = v + u, \|v\|_2 < \alpha/2, u \in \ker U\}$. For a vector $w = v + u$, $v \perp \ker U$, $u \in \ker U$, and minimum positive eigenvalue λ_+ of U,

$$
w^T U w \leq \frac{\alpha^2\lambda_+}{4} \implies v^T U v \leq \frac{\alpha^2\lambda_+}{4} \implies \|v\|_2 \leq \frac{\alpha}{2},
$$

and so $C(w)$ is continuous, has derivatives of all orders, and derivatives can be passed through the integral within the set $\{w \in \mathbb{R}^N : w^T U w \leq \frac{\alpha^2\lambda_+}{4}\}$. By Taylor's theorem, for any w in this set, there exists a distribution ν_w with density proportional to $\pi\exp(\bar{\ell}_{w'})$ for some w' on the segment from the origin to w such that

$$
C(w) = \log\int \pi\exp(\bar{\ell}_w) = \frac{1}{2}w^T U w + \frac{1}{6}\mathbb{E}_{\nu_w}\Big[(w^T(\bar{\ell}_n)_{n=1}^N)^3\Big].
$$

By definition of ν_w, $w \in \ker U$ implies that $w^T(\bar{\ell}_n)_{n=1}^N = 0\ \nu_w$-a.s. and hence $\frac{1}{6}\mathbb{E}_{\nu_w}\big[(w^T(\bar{\ell}_n)_{n=1}^N)^3\big] = 0$. Therefore, for $w^T U w \leq \frac{\alpha^2\lambda_+}{4}$,

$$
\begin{aligned}
C(w) &\leq \frac{1}{2}w^T U w\left(1 + \max_{\substack{w^T U w \leq \frac{\alpha^2\lambda_+}{4} \\ w \perp \ker U}} \frac{1}{6}\frac{\mathbb{E}_{\nu_w}\big[(w^T(\bar{\ell}_n)_{n=1}^N)^3\big]}{w^T U w}\right) \\
&\leq \frac{1}{2}w^T U w\left(1 + \max_{\|w\|_2 \leq \frac{\alpha}{2}} \frac{1}{6}\frac{\|w\|_2\big\|\mathbb{E}_{\nu_w}\big[(\bar{\ell}_n)_{n=1}^N \otimes (\bar{\ell}_n)_{n=1}^N \otimes (\bar{\ell}_n)_{n=1}^N\big]\big\|_2}{\lambda_+}\right) \\
&\leq \frac{1}{2}w^T U w\left(1 + \frac{\alpha}{12\lambda_+}\max_{\|w\|_2 \leq \frac{\alpha}{2}}\Big\|\mathbb{E}_{\nu_w}\Big[(\bar{\ell}_n)_{n=1}^N \otimes (\bar{\ell}_n)_{n=1}^N \otimes (\bar{\ell}_n)_{n=1}^N\Big]\Big\|\right),
\end{aligned}
$$

where $\otimes$ denotes outer products to form a tensor. By continuity of derivatives of all orders within the neighbourhood $\|w\|_2 < \alpha$, the result follows by selecting a sufficiently small α. □

Proposition A.1. *Suppose there exist $c \in \mathbb{R}$, $\alpha, \delta > 0$, and $0 < \epsilon < 1$ such that $\ell \le c$ and for all coreset weights w satisfying $\alpha w^T \operatorname{Cov}_\pi((\ell_n)_{n=1}^N) w \le 1$, $|\bar{\ell}_w| \le \epsilon|\ell - c| + \delta$. Then the potentials $(\ell_n)_{n=1}^N$ are $(h, \alpha \operatorname{Cov}_\pi((\ell_n)_{n=1}^N))$-subexponential with $h(x) = \frac{1}{2}x + \frac{e^{\delta + c\epsilon}}{\int \pi_0 e^{\epsilon\ell}} x^{1-\epsilon}$.*

Proof of Proposition A.1. Let $\ell' = \ell - c$. Since $\ell' \le 0$ and $|\bar{\ell}_w| \le \epsilon|\ell'| + \delta$ for some $\epsilon < 1$, $\delta > 0$,

$$\begin{aligned}
\int \pi \exp(\bar{\ell}_w) &= 1 + \frac{1}{2}\int \pi(\bar{\ell}_w)^2 + \int \pi \sum_{k=3}^{\infty} \frac{1}{k!}(\bar{\ell}_w)^{k-2(1-\epsilon)}(\bar{\ell}_w)^{2(1-\epsilon)} \\
&\le 1 + \frac{1}{2}\int \pi(\bar{\ell}_w)^2 + \int \pi \sum_{k=3}^{\infty} \frac{1}{k!}(\epsilon|\ell'| + \delta)^{k-2(1-\epsilon)}|\bar{\ell}_w|^{2(1-\epsilon)} \\
&= 1 + \frac{1}{2}\int \pi(\bar{\ell}_w)^2 + \int \pi \left(\frac{e^{\epsilon|\ell'|+\delta} - 1 - (\epsilon|\ell'| + \delta) - \frac{1}{2}(\epsilon|\ell'| + \delta)^2}{(\epsilon|\ell'| + \delta)^{2(1-\epsilon)}} \right) |\bar{\ell}_w|^{2(1-\epsilon)} \\
&\le 1 + \frac{1}{2}\int \pi(\bar{\ell}_w)^2 + \int \pi e^{\epsilon|\ell'|+\delta}|\bar{\ell}_w|^{2(1-\epsilon)} \\
&= 1 + \frac{1}{2}\int \pi(\bar{\ell}_w)^2 + \frac{\int \pi_0 e^{(1-\epsilon)\ell' + \delta}|\bar{\ell}_w|^{2(1-\epsilon)}}{\int \pi_0 e^{\ell'}} \\
&\le 1 + \frac{1}{2}\int \pi(\bar{\ell}_w)^2 + e^{\delta}\frac{\left(\int \pi_0 e^{\ell'}|\bar{\ell}_w|^2\right)^{1-\epsilon}}{\int \pi_0 e^{\ell'}} \\
&= 1 + \frac{1}{2}\int \pi(\bar{\ell}_w)^2 + e^{\delta}\left(\int \pi_0 e^{\ell'}\right)^{-\epsilon}\left(\int \pi(\bar{\ell}_w)^2\right)^{1-\epsilon} \\
&= 1 + \frac{1}{2}\int \pi(\bar{\ell}_w)^2 + e^{\delta + c\epsilon}\left(\int \pi_0 e^{\ell}\right)^{-\epsilon}\left(\int \pi(\bar{\ell}_w)^2\right)^{1-\epsilon} \\
&\le \exp\Big(h(w^T \operatorname{Cov}_\pi(\ell) w)\Big),
\end{aligned}$$

where $h(x) = \frac{1}{2}x + \frac{e^{\delta + c\epsilon}}{\int \pi_0 e^{\epsilon\ell}} x^{1-\epsilon}$. □

Proposition A.2. *Suppose $\Theta = \mathbb{R}^d$, $\bar{\ell}$ is G-strongly concave, and there exists $L < G$, $\alpha > 0$, and $\theta_0 \in \Theta$ such that for all coreset weights w satisfying $\alpha w^T \operatorname{Cov}_\pi((\ell_n)_{n=1}^N) w \le 1$, $\bar{\ell}_w$ is L-Lipschitz smooth, and both $\|\nabla \ell_w(\theta_0)\|$ and $\bar{\ell}_w(\theta_0)$ are bounded. Then for any $(L/G) < \epsilon < 1$, there exists $c \in \mathbb{R}$, $\delta > 0$ such that the potentials $(\ell_n)_{n=1}^N$ are $(h, \alpha \operatorname{Cov}_\pi((\ell_n)_{n=1}^N))$-subexponential with the same h as in Proposition A.1.*

Proof of Proposition A.2. Since $\bar{\ell}$ is G-strongly concave and $\bar{\ell}_w$ is L-Lipschitz smooth, we can write

$$\begin{aligned}
\ell(\theta) &\le \ell(\theta_0) + \nabla \ell(\theta_0)^T(\theta - \theta_0) - \frac{G}{2}\|\theta - \theta_0\|^2 \\
&= \ell(\theta_0) + \frac{G}{2}\left\|G^{-1}\nabla \ell(\theta_0)\right\|^2 - \frac{G}{2}\left\|\theta - \theta_0 - G^{-1}\nabla \ell(\theta_0)\right\|^2 \\
|\bar{\ell}_w(\theta)| &\le |\bar{\ell}_w(\theta_0) + \nabla \ell_w(\theta_0)^T(\theta - \theta_0)| + \frac{L}{2}\|\theta - \theta_0\|^2.
\end{aligned}$$

So setting $c = \ell(\theta_0) + \frac{G}{2}\left\|G^{-1}\nabla \ell(\theta_0)\right\|^2$ implies $\ell - c$ is a nonpositive function as required. Then

$$|\bar{\ell}_w(\theta)| - \epsilon|\ell(\theta) - c| \le |\bar{\ell}_w(\theta_0)| + \frac{\epsilon}{2G}\|\nabla \ell(\theta_0)\|^2 + (\|\nabla \ell_w(\theta_0)\| + \epsilon\|\nabla \ell(\theta_0)\|)\|\theta - \theta_0\| + \frac{L - \epsilon G}{2}\|\theta - \theta_0\|^2.$$

For $0 < a < G - L$, setting $\epsilon = \frac{L + a}{G}$ and then maximizing over $\|\theta - \theta_0\|$ yields

$$\begin{aligned}
|\bar{\ell}_w(\theta)| - \epsilon|\ell(\theta) - c| &\le |\bar{\ell}_w(\theta_0)| + \frac{\epsilon}{2G}\|\nabla \ell(\theta_0)\|^2 + (\|\nabla \ell_w(\theta_0)\| + \epsilon\|\nabla \ell(\theta_0)\|)\|\theta - \theta_0\| - \frac{a}{2}\|\theta - \theta_0\|^2 \\
&\le |\bar{\ell}_w(\theta_0)| + \frac{\epsilon}{2G}\|\nabla \ell(\theta_0)\|^2 + \frac{(\|\nabla \ell_w(\theta_0)\| + \epsilon\|\nabla \ell(\theta_0)\|)^2}{2a}.
\end{aligned}$$

By the boundedness of $\bar{\ell}_w(\theta_0)$ and $\nabla \ell_w(\theta_0)$, maximizing over w yields a value of $\delta < \infty$. □

Lemma A.3. *Let $X_1, X_2, \ldots$ be i.i.d. random variables in $\mathbb{R}$ with $\mathbb{E}X_n = 0$, and define the resampled sum*

$$S_N = \sum_{n=1}^{N} \frac{M_n}{M p_n} X_n$$

where $(M_1, \ldots, M_N) \sim \text{Multi}(M, (p_1, \ldots, p_N))$*, with strictly positive resampling probabilities* $p_1, \ldots, p_N$ *that may depend on* $X_1, \ldots, X_N$ *and* N*. If there exists a* $\delta > 0$ *such that as* $N \to \infty$*,*

$$\frac{1}{N}\sum_n \frac{|X_n|^{2+\delta}}{(Np_n)^{1+\delta}} = O_p(1), \quad \frac{1}{N}\sum_n \frac{X_n^2}{Np_n} = \Omega_p(1), \quad \textit{and} \quad M \to \infty,$$

then

$$\sqrt{M}\frac{\frac{1}{N}S_N - \frac{1}{N}\sum_n X_n}{\sqrt{\frac{1}{N}\sum_n \frac{X_n^2}{Np_n}}} \xrightarrow{d} \mathcal{N}(0,1).$$

Proof. We can rewrite

$$S_N = \frac{1}{M}\sum_{m=1}^M \frac{X_{I_m}}{p_{I_m}}$$

where $I_m \overset{\text{iid}}{\sim} \text{Categorical}(p_1, \ldots, p_N)$. Consider $S_N + B_N$ where B_N is independent of S_N, $B_N = \pm 1$ with probability $\frac{1}{2(NM)^{1+\delta}}$, and $B_N = 0$ otherwise. So if we set $\mathcal{A}_N = \sigma(X_1, \ldots, X_N)$, [54, Cor. 3] asserts that

$$\frac{S_N + B_N - \mathbb{E}[S_N|\mathcal{A}_N]}{\sqrt{(NM)^{-(1+\delta)} + \text{Var}[S_N|\mathcal{A}_N]}} \xrightarrow{d} \mathcal{N}(0,1) \qquad N \to \infty.$$

as long as for all N large enough,

$$\text{Var}\left[\frac{1}{M}\frac{X_{I_m}}{p_{I_m}}|\mathcal{A}_N\right] < \infty \quad \text{a.s.},$$

and as $N \to \infty$,

$$\frac{(NM)^{-(1+\delta)} + \sum_{m=1}^M \mathbb{E}\left[\left|\frac{1}{M}\frac{X_{I_m}}{p_{I_m}} - \mathbb{E}\left[\frac{1}{M}\frac{X_{I_m}}{p_{I_m}}|\mathcal{A}_N\right]\right|^{2+\delta}|\mathcal{A}_N\right]}{((NM)^{-(1+\delta)} + \text{Var}[S_N|\mathcal{A}_N])^{(2+\delta)/2}} \xrightarrow{p} 0.$$

Note that the conditional mean and variance have the form

$$\mathbb{E}[S_N|\mathcal{A}_N] = \mathbb{E}\left[\frac{X_{I_m}}{p_{I_m}}|\mathcal{A}_N\right] = \sum_n X_n$$

$$\text{Var}[S_N|\mathcal{A}_N] = \frac{1}{M}\text{Var}\left[\frac{X_{I_m}}{p_{I_m}}|\mathcal{A}_N\right] = \frac{1}{M}\sum_n p_n\left(\frac{X_n}{p_n} - \sum_n X_n\right)^2,$$

which implies that $\text{Var}\left[\frac{1}{M}\frac{X_{I_m}}{p_{I_m}}|\mathcal{A}_N\right] < \infty$ a.s., since $p_1, \ldots, p_N$ are strictly nonnegative and $\mathbb{E}X_n = 0$ implies X_n is finite almost surely. Next, note that

$$\begin{aligned}
&\frac{(NM)^{-(1+\delta)} + \sum_{m=1}^M \mathbb{E}\left[\left|\frac{1}{M}\frac{X_{I_m}}{p_{I_m}} - \mathbb{E}\left[\frac{1}{M}\frac{X_{I_m}}{p_{I_m}}|\mathcal{A}_N\right]\right|^{2+\delta}|\mathcal{A}_N\right]}{((NM)^{-(1+\delta)} + \text{Var}[S_N|\mathcal{A}_N])^{(2+\delta)/2}} \\
\leq &\frac{(NM)^{-(1+\delta)} + 2^{2+\delta}\sum_{m=1}^M \left(\mathbb{E}\left[\left|\frac{1}{M}\frac{X_{I_m}}{p_{I_m}}\right|^{2+\delta}|\mathcal{A}_N\right] + \left|\frac{1}{M}\sum_n X_n\right|^{2+\delta}\right)}{((NM)^{-(1+\delta)} + \text{Var}[S_N|\mathcal{A}_N])^{(2+\delta)/2}} \\
= &M^{-\delta/2}\frac{N^{-(3+2\delta)} + 2^{2+\delta}\left(\frac{1}{N}\sum_n \frac{|X_n|^{2+\delta}}{(Np_n)^{1+\delta}} + \left|\frac{1}{N}\sum_n X_n\right|^{2+\delta}\right)}{\left(M^{-\delta}N^{-(3+\delta)} + \frac{1}{N}\sum_n \frac{X_n^2}{Np_n} - \left(\frac{1}{N}\sum_n X_n\right)^2\right)^{(2+\delta)/2}}.
\end{aligned}$$

The above expression converges in probability to 0 by the technical assumptions in the statement of the result as well as the fact that $\frac{1}{N}\sum_n X_n \xrightarrow{a.s.} 0$ by the law of large numbers. Once again by the technical assumptions, $\text{Var}[S_N|\mathcal{A}_N] = \Omega_p(N^2/M)$, so

$$\frac{\text{Var}[S_N|\mathcal{A}_N]}{(NM)^{-(1+\delta)} + \text{Var}[S_N|\mathcal{A}_N]} \xrightarrow{p} 1$$

$$\frac{B_N}{(NM)^{-(1+\delta)} + \text{Var}[S_N|\mathcal{A}_N]} \xrightarrow{p} 0,$$

and hence by Slutsky's theorem,

$$\frac{S_N - \mathbb{E}[S_N|\mathcal{A}_N]}{\sqrt{\text{Var}[S_N|\mathcal{A}_N]}} \xrightarrow{d} \mathcal{N}(0,1) \qquad N \to \infty.$$

Using Slutsky's theorem again with $\frac{1}{N}\sum_n X_n \xrightarrow{p} 0$ and rearranging yields the final result. □

Lemma A.4. *Suppose coreset weights are generated using the importance weighted construction in Algorithm 1. Let* $g = \nabla \ell(\eta_0)$, $g_w = \nabla \ell_w(\eta_0)$, *and* $H = -\mathbb{E}\big[\nabla^2 \ell_n(\eta_0)\big]$. *If conditions A(1-3) and (A6) in Section 4 hold,* $M = o(N)$, *and* $M = \omega(1)$, *then*

$$\left\| \frac{g}{N} \right\|_2 = \Theta_p\Big(N^{-1/2}\Big), \qquad \left\| \frac{g_w}{\overline{w}} \right\|_2 = \Theta_p\left(M^{-1/2}\right), \qquad \frac{\overline{w}}{N} \xrightarrow{p} 1,$$

and

$$\sup_{\|\eta-\eta_0\|_2 \le r} \left\| -\frac{1}{N}\nabla^2 \ell(\eta) - H \right\|_2 \xrightarrow{p} 0, \qquad \sup_{\|\eta-\eta_0\|_2 \le r} \left\| -\frac{1}{\overline{w}}\nabla^2 \ell_w(\eta) - H \right\|_2 \xrightarrow{p} 0.$$

Proof. First, since $\overline{w} = \sum_n \frac{M_n}{Mp_n}$, $\mathbb{E}\overline{w} = N$, and

$$\begin{aligned}
\mathbb{E}\big[(\overline{w} - N)^2\big] &= \frac{N^2}{M^2}\mathbb{E}\left[\left(\sum_n M_n\big((Np_n)^{-1} - 1\big)\right)^2\right] \\
&= \frac{N^2}{M^2}\left(\sum_n ((Np_n)^{-1} - 1)^2 \mathbb{E}M_n^2 + \sum_{n \ne n'} ((Np_n)^{-1} - 1)((Np_{n'})^{-1} - 1)\mathbb{E}[M_n M_{n'}]\right) \\
&= \frac{N^2}{M}\left(\sum_n ((Np_n)^{-1} - 1)^2 p_n - \left(\sum_{n'} (1/N - p_n)\right)^2\right) \\
&= \frac{1}{M}\left(\sum_n p_n (p_n^{-1} - N)^2\right) \\
&\le \frac{1}{M}\Big(\max_n (p_n^{-1} - N)^2\Big) \\
&\le \frac{N^2}{M} O(1),
\end{aligned}$$

where the last line follows by assumption A6. Therefore by Chebyshev's inequality and $M \to \infty$, $\overline{w}/N \xrightarrow{p} 1$. Since the data are i.i.d., by conditions A1 and A2, the central limit theorem holds for the sum of $\nabla \ell_n(\eta_0)$ such that $g/\sqrt{N}$ converges in distribution to a normal, and hence $\left\| \frac{g}{N} \right\| = \Theta_p(N^{-1/2})$. By conditions A1, A2, and A6, Lemma A.3 holds such that for any $t \in \mathbb{R}^d$,

$$\sqrt{M}\frac{\frac{1}{N}t^T g_w - \frac{1}{N}t^T g}{\sqrt{\frac{1}{N}\sum_n \frac{(t^T \nabla \ell_n(\eta_0))^2}{Np_n}}} = \Theta_p(1).$$

Since condition A6 asserts that $C > Np_n \ge c > 0$, the law of large numbers, condition A1, and $M/N \to 0$ imply that

$$\frac{\sqrt{M}}{N} t^T g_w = \Theta_p(1).$$

Summing over a basis of vectors $t_1, \ldots, t_d$ shows that

$$\frac{\sqrt{M}}{N}\|g_w\|_2 = \Theta(1)\sqrt{M}\left\|\frac{g_w}{\overline{w}}\right\|_2 \Theta_p(1).$$

This completes the first three results. Next, by condition A3, for sufficiently large N such that the neighbourhood contains the ball of radius r around η_0,

$$\begin{aligned}
\sup_{\|\eta-\eta_0\|_2 \le r} \left\| \frac{1}{N}\nabla^2 \ell(\eta) - \frac{1}{N}\nabla^2 \ell(\eta_0) \right\|_2 &\le r\frac{1}{N}\sum_n R(X_n) \\
\sup_{\|\eta-\eta_0\|_2 \le r} \left\| \frac{1}{N}\nabla^2 \ell_w(\eta) - \frac{1}{N}\nabla^2 \ell_w(\eta_0) \right\|_2 &\le r\frac{1}{N}\sum_n w_n R(X_n),
\end{aligned}$$

and

$$\mathbb{E}\left[r\frac{1}{N}\sum_n R(X_n)\right] = \mathbb{E}\left[r\frac{1}{N}\sum_n w_n R(X_n)\right] = r\mathbb{E}\left[R(X)\right] \to 0,$$

so that we have that both

$$\sup_{\|\eta-\eta_0\|_2 \le r} \left\| \frac{1}{N}\nabla^2 \ell(\eta) - \frac{1}{N}\nabla^2 \ell(\eta_0) \right\|_2 \xrightarrow{p} 0 \quad \text{and} \quad \sup_{\|\eta-\eta_0\|_2 \le r} \left\| \frac{1}{N}\nabla^2 \ell_w(\eta) - \frac{1}{N}\nabla^2 \ell_w(\eta_0) \right\|_2 \xrightarrow{p} 0$$

by Markov's inequality. Finally, by the bounded variance in A2, sampling probability bounds in A6, and $M \to \infty$, the variances of $\frac{1}{N}\nabla^2 \ell_w(\eta_0)$ and $\frac{1}{N}\nabla^2 \ell(\eta_0)$ both converge to 0 as $N \to \infty$, and since both of these quantities are unbiased estimates of $\mathbb{E}\big[\nabla^2 \ell_n(\eta_0)\big]$, Chebyshev's inequality yields the desired convergence in probability. □

Lemma A.5. *Suppose $(X_n)_{n=1}^N$ are N i.i.d. random vectors in $\mathbb{R}^d$. Fix $M \in \mathbb{N}$, $M < d$ and define $X = \begin{bmatrix} X_1 & X_2 & \dots & X_M \end{bmatrix} \in \mathbb{R}^{d \times M}$. If there exists $\delta > 0$ such that*

$$\mathbb{E}\Big[(1^T (X^T X)^{-1} 1)^{M+\delta}\Big] < \infty,$$

where 1 denotes a vector of all 1 entries, then as $N \to \infty$,

$$\begin{pmatrix} \min_{w \in \mathbb{R}_+^N} \left\| \frac{\sum_{n=1}^N w_n X_n}{\sum_{n=1}^N w_n} \right\|^2 \\ \text{s.t. } \sum_n \mathbb{1}[w_n > 0] < M. \end{pmatrix} = \omega_p\Big(N^{-\frac{M+\delta/2}{M+\delta}}\Big).$$

Proof. For any $\epsilon > 0$, by the union bound over subsets of $[N]$ of size M,

$$\begin{aligned}
\mathbb{P}\left(\min_{w \in \mathbb{R}_+^N} \cdots \le \epsilon\right) &\le \binom{N}{M} \mathbb{P}\left(\min_{w \in \mathbb{R}^M} \frac{w^T X^T X w}{w^T 1 1^T w} \le \epsilon\right) \\
&\le \binom{N}{M} \mathbb{P}\left(\max_{\lambda} \min_{w \in \mathbb{R}^M} w^T X^T X w - \lambda(1^T w - 1) \le \epsilon\right) \\
&= \binom{N}{M} \mathbb{P}\left(\max_{\lambda} \lambda - \frac{\lambda^2}{4} 1^T (X^T X)^{-1} 1 \le \epsilon\right) \\
&= \binom{N}{M} \mathbb{P}\Big(1^T (X^T X)^{-1} 1 \ge \epsilon^{-1}\Big).
\end{aligned}$$

By Markov's inequality and $\binom{N}{M} \le (eN/M)^M$,

$$\begin{aligned}
\mathbb{P}\left(\min_{w \in \mathbb{R}_+^N} \cdots \le \epsilon\right) &\le \left(\frac{eN}{M}\right)^M \epsilon^{M+\delta} \mathbb{E}\Big[(1^T (X^T X)^{-1} 1)^{M+\delta}\Big] \\
&= \left(\frac{eN\epsilon^{\frac{M+\delta}{M}}}{M}\right)^M \mathbb{E}\Big[(1^T (X^T X)^{-1} 1)^{M+\delta}\Big].
\end{aligned}$$

Setting $\epsilon = N^{-\frac{M+\delta/2}{M+\delta}}$ yields

$$\mathbb{P}\left(\min_{w \in \mathbb{R}_+^N} \cdots \le N^{-\frac{M+\delta/2}{M+\delta}}\right) \le \left(\frac{eN^{-\frac{\delta}{2M}}}{M}\right)^M \mathbb{E}\Big[(1^T (X^T X)^{-1} 1)^{M+\delta}\Big].$$

The right-hand side converges to 0 as $N \to \infty$, yielding the stated result. □

Proof of Corollary 4.1 and Corollary 4.2. Set $r = (\log M)^{-1/2}$. Then since $M = o(N)$, $M = \omega(1)$, and assumptions (A1-3) and (A6) hold, Lemma A.4 holds. Note that $\|g_w/\overline{w}\| = \Theta_p(M^{-1/2}) = o_p(r)$, $\eta\pi_0$ is positive at η_0 and twice differentiable by (A4), and $Nr^2 = N/\log M = \omega(1)$ since $M = o(N)$. Thus the conditions of Theorem 3.3 are verified. Substitution into the right term in the minimum of Theorem 3.3 yields the stated lower bound of $\Omega_p(N/M)$. For the left term in Theorem 3.3, define $B = \{(\eta - \eta_0)^T H (\eta - \eta_0) \le r^2\}$. Then since $H \succ 0$, $r \to 0$, and $r^2 = 1/\log M = \omega(\log N/N)$, (A5) guarantees that $-\log(\eta\pi)(B^c) = \Omega_p(Nr^2) = \Omega_p(N/\log M)$. Therefore the minimum is $\Omega_p(N/M)$, and we complete the proof by transferring from $\underline{\mathrm{KL}}(w)$ on the η-pushforward model to $\underline{\mathrm{KL}}(w)$ on the original model using Corollary 3.6. □

Proof of Corollary 4.3. Fix the $\delta > 0$ guaranteed by (A8), and set $r = N^{-\frac{M+3\delta/4}{2(M+\delta)}}$. Note that $Nr^2 = N^{\frac{\delta/4}{M+\delta}} = \omega(1)$, $\eta\pi_0$ is positive at η_0 and twice differentiable by (A4), and by (A1-3) the results pertaining to $\left\|\frac{g}{N}\right\|_2$ and $\sup_{\|\eta-\eta_0\|_2 \le r} \left\| -\frac{1}{N}\nabla^2 \ell(\eta) - H \right\|_2$ in Lemma A.4 hold; thus Assumption 3.2 holds. By (A7), Assumption 3.4 holds as well as the conditions on $\frac{1}{\overline{w}}\nabla^2 \ell_w(\theta)$ and $\frac{1}{\overline{w}}\sum_n w_n L_n^2$ in Theorem 3.5. Finally by (A8), Lemma A.5 holds such that

$$\left\|\frac{g_w}{\overline{w}}\right\|^2 = \omega_p\Big(N^{-\frac{M+\delta/2}{M+\delta}}\Big),$$

and hence $\left\|\frac{g_w}{\overline{w}}\right\| = \omega_p(r)$. Therefore all conditions of Theorem 3.5 hold. For the left term in the minimum in Theorem 3.5, define $B = \{(\eta-\eta_0)^T H(\eta-\eta_0) \le r^2\}$. Then since $H \succ 0$, $r \to 0$, and $r^2 = N^{-\frac{M+3\delta/4}{M+\delta}} = \omega(\log N/N)$, (A5) guarantees that $-\log(\eta\pi)(B^c) = \Omega_p(Nr^2) = \Omega_p\left(N^{\frac{\delta/4}{M+\delta}}\right)$. For the right term,

$$\log\left(N\left\|\frac{g_w}{\overline{w}}\right\|^2\right) = \Omega_p\left(\log N^{1-\frac{M+\delta/2}{M+\delta}}\right) = \Omega_p(\log N).$$

The minimum of these two is from the right term, so

$$\underline{\mathrm{KL}}(w) = \Omega_p(\log N).$$

We complete the proof by transferring from $\underline{\mathrm{KL}}(w)$ on the η-pushforward model to $\underline{\mathrm{KL}}(w)$ on the original model using Corollary 3.6. □

Proposition A.6. *The models specified in Eqs. (2) and (3) satisfy assumptions (A1-5).*

Proof. The exact same technique applies to both models, so here we will just demonstrate it for the Cauchy model. In the Cauchy model, $\theta \in \mathbb{R}$, $\eta : \mathbb{R} \to \mathbb{R}_+$, $\eta(\theta) = \theta^2$, and

$$\ell_n(\eta) = -\log\pi - \log((Z_n-\eta)^2+1) \qquad \nabla\ell_n(\eta) = \frac{2(Z_n-\eta)}{(Z_n-\eta)^2+1}$$
$$\nabla^2\ell_n(\eta) = \frac{2(Z_n-\eta)^2-2}{((Z_n-\eta)^2+1)^2} \qquad \nabla^3\ell_n(\eta) = \frac{4((Z_n-\eta)^2-3)(\eta-Z_n)}{((\eta-Z_n)^2+1)^3},$$

where $Z_n = X_n$. Property (A1) holds by routine interchange of differentiation and integration. Property (A2) holds (for any $\delta > 0$) because $\nabla\ell_n(\eta)$ and $\nabla^2\ell_n(\eta)$ are bounded functions of η and X_n jointly. Property (A3) holds (for any neighbourhood of η_0) because $\nabla^3\ell_n(\eta)$ is a bounded function of η and X_n jointly. Property (A4) holds because the pushforward of $\mathrm{Cauchy}(0,1)$ through $\eta(\theta) = \theta^2$ has full support on $\mathbb{R}_+$. In order to verify assumption (A5), suppose there exists a sequence of bounded measurable functions $\phi_r(Z_1,\dots,Z_N) \in [0,1]$ of the data and constants $c, c' > 0$ such that for all $r \to 0$, $r^2 = \omega(\log N/N)$,

$$\mathbb{E}_{\eta_0}\phi_r = O\left(e^{-cNr^2}\right) \quad \text{and} \quad \sup_{\|\eta-\eta_0\|>r} \mathbb{E}_\eta(1-\phi_r) = O\left(e^{-c'Nr^2}\right).$$

The functions ϕ_r are similar to the test functions of Schwartz [55]. Then defining $\mu = \eta\pi$ and $\mu_0 = \eta\pi_0$,

$$\begin{aligned}\mu(\|\eta-\eta_0\|>r) &= \phi_r\mu(\|\eta-\eta_0\|>r) + (1-\phi_r)\mu(\|\eta-\eta_0\|>r)\\ &\le \phi_r + (1-\phi_r)\mu(\|\eta-\eta_0\|>r)\\ &= \phi_r + \frac{\int_{\|\eta-\eta_0\|>r}(1-\phi_r)e^{\ell(\eta)-\ell(\eta_0)}\mu_0}{\int e^{\ell(\eta)-\ell(\eta_0)}\mu_0}.\end{aligned}$$

Using the same proof technique as in Theorem 3.3, the denominator satisfies

$$\log\int e^{\ell(\eta)-\ell(\eta_0)}\mu_0(\mathrm{d}\eta) \ge -\frac{d}{2}\log N + O_p(1).$$

By assumption, there exists $c > 0$ such that

$$\mathbb{E}_{\eta_0}[\phi_r] = O\left(e^{-cNr^2}\right) \implies \phi_r = O_p(e^{-cNr^2}),$$

and a $c' > 0$ such that

$$\begin{aligned}\mathbb{E}_{\eta_0}\left[\int_{\|\eta-\eta_0\|>r}(1-\phi_r)e^{\ell(\eta)-\ell(\eta_0)}\mu_0\right] &= \int_{\|\eta-\eta_0\|>r}\mathbb{E}_\eta(1-\phi_r)\mu_0\\ &\le \sup_{\|\eta-\eta_0\|>r}\mathbb{E}_\eta(1-\phi_r)\\ &= O(e^{-cNr^2}) \implies \int_{\|\eta-\eta_0\|>r}(\dots) = O_p\left(e^{-cNr^2}\right).\end{aligned}$$

Therefore $\mu(\|\eta-\eta_0\| \ge r) = O_p\left(e^{-cNr^2} + N^{d/2}e^{-c'Nr^2}\right) = O_p\left(e^{(d/2)\log N - c''Nr^2}\right)$; and since $r^2 = \omega(\log N/N)$, $-\log\mu(\|\eta-\eta_0\| \ge r) = \Omega_p(Nr^2)$ as required by (A5). So to complete the proof of (A5) we need to find a suitable ϕ_r. Fix $\epsilon > 0$, and set

$$\phi_r(Z_1,\dots,Z_N) = \mathbb{1}\left[P_{\eta_0}(|Z-\eta_0| \le 1) - \frac{1}{N}\sum_{n=1}^N \mathbb{1}[|Z_n-\eta_0| \le 1] > \epsilon r\right].$$

Under p_{η_0}, Hoeffding's inequality yields

$$\mathbb{E}_{\eta_0}\phi_r \le e^{-2N\epsilon^2 r^2}.$$

And under p_η for $\|\eta - \eta_0\| > r$, for small enough $\epsilon > 0$, $P_{\eta_0}(|Z - \eta_0| \le 1) - P_\eta(|Z - \eta_0| \le 1) \ge 2\epsilon r$. Therefore

$$\begin{aligned}
\mathbb{E}_\eta[1 - \phi_r(Z_1, \dots, Z_N)] &= \Pr_\eta\left(P_{\eta_0}(|Z - \eta_0| \le 1) - \frac{1}{N}\sum_{n=1}^N \mathbb{1}[|Z_n - \eta_0| \le 1] \le \epsilon r\right) \\
&\le \Pr_\eta\left(P_\eta(|Z - \eta_0| \le 1) - \frac{1}{N}\sum_{n=1}^N \mathbb{1}[|Z_n - \eta_0| \le 1] \le -\epsilon r\right) \\
&= \Pr_\eta\left(\frac{1}{N}\sum_{n=1}^N \mathbb{1}[|Z_n - \eta_0| \le 1] - P_\eta(|Z - \eta_0| \le 1) \ge \epsilon r\right),
\end{aligned}$$

at which point we can again apply Hoeffding's inequality, completing the result. □

Lemma A.7. *Fix vectors $u, u_1, \dots, u_N$ in a separable Hilbert space with inner product denoted $a \cdot b$ and norm denoted $\|\ \|$. Let $v_1, \dots, v_M$ be drawn from $\{u_1, \dots, u_N\}$ with probabilities $p_1, \dots, p_N$ either with or without replacement (if without replacement, the probabilities are renormalized after every draw). Then for all $\epsilon \ge 0$,*

$$\mathbb{P}\left(\min_{w \ge 0}\left\|\sum_{m=1}^M w_m v_m - u\right\|^2 > \epsilon^{M\left(\frac{q(M,\epsilon)}{2}\right)+1}\|u\|^2\right) \le e^{-\left(\frac{1-\log(2)}{2}\right)M},$$

where

$$q(M,\epsilon) = \mathbb{P}\left(1 - \max\left\{0, \frac{v_M}{\|v_M\|} \cdot \frac{(u - x_{M-1})}{\|u - x_{M-1}\|}\right\}^2 \le \epsilon\right) \quad x_{M-1} = \operatorname*{arg\,min}_{x \in \operatorname{cone}\{v_1, \dots, v_{M-1}\}} \|x - u\|^2.$$

Proof. First note that it suffices to analyze the case with replacement, since this case provides an upper bound on the case without replacement. To demonstrate this, we couple two probability spaces—one that draws $v_1, \dots, v_M$ with replacement, and one without replacement. First, draw an identical vector v_1 for both copies. On each subsequent iteration $m > 1$, the "with replacement" copy first draws whether or not it selects a vector that was previously selected by the "without replacement" copy. If it does, it draws that vector independently; if it does not, it selects the same vector as the "without replacement" copy. In any case, at each iteration m, the vectors drawn by the "with replacement" copy are always a subset of the vectors drawn by the "without replacement" copy, and hence the minimum over $w \ge 0$ is greater for that copy. It therefore suffices to analyze the case with replacement.

To obtain an upper bound on the probability when sampling with replacement, instead of minimizing over all $w \ge 0$ jointly, suppose we use the following iterative algorithm. Set $x_0 = 0$. At the first iteration, we draw v_1 and set the weight w_1 by optimizing over $w_1 \ge 0$:

$$\min_{w_1 > 0}\|w_1 v_1 - u\|^2 = \|u\|^2\left(1 - \max\left\{0, \frac{v_1 \cdot u}{\|v_1\|\|u\|}\right\}^2\right).$$

Set $x_1 = w_1 v_1$, and note that $(u - x_1) \cdot x_1 = 0$. Then at each subsequent iteration k, assume the previous iterate is optimized over all nonnegative weights, and hence satisfies $(u - x_{k-1}) \cdot x_{k-1} = 0$. We draw another vector v_k, and bound the erorr of the next iterate x_k by optimizing over only the weight w_k for the new vector v_k. Then

$$\begin{aligned}
\|u - x_k\|^2 = \min_{w_1, \dots, w_k \ge 0}\left\|\sum_{m=1}^k w_m v_m - u\right\|^2 &\le \min_{w_k > 0}\|w_k v_k + x_{k-1} - u\|^2 \\
&= \|u - x_{k-1}\|^2\left(1 - \max\left\{0, \frac{v_k \cdot (u - x_{k-1})}{\|v_k\|\|u - x_{k-1}\|}\right\}^2\right).
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\mathbb{P}\left(\min_{w \ge 0}\left\|\sum_{m=1}^M w_m v_m - u\right\|^2 \le \epsilon^K\|u\|^2\right) \\
&\ge \mathbb{P}\big(\text{in at least } K \text{ iterations, } \|x_k - u\|^2 \le \epsilon\|x_{k-1} - u\|^2\big)
\end{aligned}$$

$$
\begin{aligned}
&\geq \mathbb{P}\left(\text{in at least } K \text{ iterations, } 1 - \max\left\{0, \frac{v_k \cdot (u - x_{k-1})}{\|v_k\|\|u - x_{k-1}\|}\right\}^2 \leq \epsilon\right) \\
&= \sum_{\mathcal{K} \subseteq [M], |\mathcal{K}| \geq K} \mathbb{P}\left(k \in \mathcal{K} \iff 1 - \max\left\{0, \frac{v_k \cdot (u - x_{k-1})}{\|v_k\|\|u - x_{k-1}\|}\right\}^2 \leq \epsilon\right) \\
&\geq \sum_{\mathcal{K} \subseteq [M], |\mathcal{K}| \geq K} q^k (1-q)^{M-k} \\
&= \sum_{k=K}^{M} \binom{M}{k} q^k (1-q)^{M-k},
\end{aligned}
$$

where

$$
\begin{aligned}
q &= \mathbb{P}\left(1 - \max\left\{0, \frac{v_M \cdot (u - x_{M-1})}{\|v_M\|\|u - x_{M-1}\|}\right\}^2 \leq \epsilon\right) \\
x_{M-1} &= \operatorname*{arg\,min}_{x \in \operatorname{cone}\{v_1, \ldots, v_{M-1}\}} \|x - u\|^2
\end{aligned}
$$

So for all $0 \leq K \leq M$,

$$
\mathbb{P}\left(\min_{w \geq 0} \left\| \sum_{m=1}^{M} w_m v_m - u \right\|^2 > \epsilon^K \|u\|^2\right) \leq \operatorname{Binom}(M, K-1, q).
$$

Using the Chernoff bound on the binomial CDF, for all $K - 1 \leq Mq$,

$$
\begin{aligned}
\mathbb{P}\left(\min_{w \geq 0} \left\| \sum_{m=1}^{M} w_m v_m - u \right\|^2 > \epsilon^K \|u\|^2\right) &\leq e^{-M\left(\frac{K-1}{M} \log \frac{K-1}{Mq} + (1 - \frac{K-1}{M}) \log \frac{1 - \frac{K-1}{M}}{1-q}\right)} \\
&= e^{-(K-1) \log \frac{K-1}{Mq} - (M - (K-1)) \log \frac{M - (K-1)}{M(1-q)}}.
\end{aligned}
$$

Substituting $K - 1 = Mq/2$ yields

$$
= e^{M((q/2) \log 2 - (1 - q/2) \log \frac{1 - q/2}{(1-q)})} \leq e^{-\left(\frac{1 - \log(2)}{2}\right) M}.
$$

□

Proof of Corollary 6.1. Since the potentials are $\beta \operatorname{Cov}_\pi((\ell_n)_{n=1}^N)$ subexponential, Theorem 5.3 guarantees that

$$
\forall w \in \mathbb{R}_+^N : 4\beta(w-1)^T \operatorname{Cov}_\pi((\ell_n)_{n=1}^N)(w-1) \leq 1, \qquad \overline{\mathrm{KL}}(w) \leq 4\beta(w-1)^T \operatorname{Cov}_\pi((\ell_n)_{n=1}^N)(w-1).
$$

We apply Lemma A.7 with vectors $\ell_1, \ldots, \ell_N$ (in equivalence classes specified up to a additive constant) and inner product between ℓ_i, ℓ_j defined by $\operatorname{Cov}_\pi(\ell_i, \ell_j)$. In the notation of Lemma A.7, by assumption, $\|u\|^2 = O_p(N^\alpha)$ and $q(M, \epsilon) = \omega_p(M^{-\rho})$. Substituting $M = (\log N)^{1/(1-\rho)}$, we find that

$$
\mathbb{P}\left(4\beta(w-1)^T \operatorname{Cov}_\pi((\ell_n)_{n=1}^N)(w-1) \geq \epsilon^{-\omega_p(\log N) + \alpha \log N}\right) \to 0.
$$

Combining this result with the KL bound above yields the final result. □

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state that the paper produces new general upper and lower bounds for Bayesian coreset approximations, that these bounds are applied to particular cases of interest, and that empirical results align with the theory. The paper does indeed contain these results, with proofs in the supplemental material, and empirical results in the figures do match the theory well.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: In the conclusions section of the work, two main limitations of the theory are mentioned: the need to perform case-by-case analysis of subexponentiality constants and alignment probabilities in Definition 5.2 and Corollary 6.1.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Theoretical results are numbered, and proofs of all results are included in the appendix.

   Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: All details needed to reproduce the experimental results in Figure 2 and 3 are included in the text. Algorithms used in the experiments exist in the cited literature.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: There are no new algorithms presented in this work; the experiments involve only existing methods for which public code is available. The code is not central to the contributions of the paper.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.

- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: All details needed to reproduce the experimental results in Figure 2 and 3 are included in the text. Algorithms used in the experiments exist in the cited literature.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: All empirical results show the mean over a number of trials, as well as error bars indicating standard error.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

Justification: These details are not important for this paper, as there are no new methods or algorithms presented or claims related to computational performance. However, the introduction does list that simulations were performed on a desktop computer with a Core i7 processor and 32GB RAM.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: This paper presents a new theoretical analysis of error bounds for Bayesian coresets methods. It does not present any new methodology or data with potential harmful consequences.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [NA]

Justification: There is no potential negative societal impact of this work. The paper provides new theory regarding existing methodology.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: This does not apply.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [NA]

Justification: This does not apply.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: This does not apply.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: This does not apply.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: This does not apply.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.