Fast Bayesian Coresets via Subsampling and Quasi-Newton Refinement: Supplementary Material

Cian Naik
Department of Statistics
University of Oxford
cian.naik@stats.ox.ac.uk

Judith Rousseau
Department of Statistics
University of Oxford
judith.rousseau@stats.ox.ac.uk

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

A Theoretical analysis

In this section we give further details on the theoretical results presented in Section 4. We start by giving the exact conditions needed for Theorem 4.2 to hold.

A.1 Full conditions for Theorem 4.2

In order to state Theorem 4.2, we define θ_0 as the unique value such that $\bar{f}(\theta) \leq \bar{f}(\theta_0)\ \forall \theta \in \Theta$, where $\bar{f}(\cdot) = \mathbb{E}_p(f_n(\cdot))$. If we split $f(x;\theta)$ into two parts,

$$f(x,\theta) = f^{(1)}(x,\theta) + f^{(2)}(x,\theta),$$

then we require that there exist constants $\epsilon_0, L_0, L_1, L_2 > 0$ such that:

• **A1**: For all x, $f(x,\theta)$ is differentiable and strongly concave in θ with constant L_0.
• **A2**: For all $0 < \epsilon < \epsilon_0$ and $\|\theta - \theta_0\| \leq \epsilon$,

$$|f^{(2)}(x,\theta)| \leq R(x) r(\epsilon) \|\theta - \theta_0\|^2,$$

where $R(x), r(\epsilon)$ are positive functions, r is monotone increasing, $\lim_{\epsilon \to 0} r(\epsilon) = 0$, and

$$4r(\epsilon_0)\left[\mathbb{E}_p(R(X)) + 1\right] \leq L_0, \quad \mathbb{E}_p(R(X)^2) < \infty.$$

• **A3**: For all $\|\theta - \theta_0\| \leq \epsilon_0$,

$$\bar{f}(\theta_0) - \bar{f}(\theta) \leq L_1 \|\theta - \theta_0\|^2, \quad \mathrm{Var}_p(f(X;\theta_0) - f(X;\theta)) \leq L_2 \|\theta - \theta_0\|^2.$$

• **A4**: $\inf_{\|\theta - \theta_0\| \leq \epsilon_0} \pi_0(\theta) > 0$, and $\sup_\theta \pi_0(\theta) < \infty$.

Here, the functions $R(x), r(\epsilon)$ control the size of the "discarded part" $f^{(2)}$ locally around θ_0; these should be small to ensure that $f(x,\theta) \approx f^{(1)}(x,\theta)$ locally. Note also that when $r(\epsilon) \asymp \epsilon$ as $\epsilon \to 0$, the result of Theorem 4.2 holds even if D is increasing such that $D = o(\log N)$.

A.2 Discussion on conditions of Theorems 4.1 to 4.3

In order for Theorem 4.1 to hold, we require that S_0 is finite dimensional (say with dimension d), where S_0 is the vector space of functions on Θ spanned by $\{f(x,\cdot); x \in \mathcal{X}\}$. For this to hold, we need to find a set of d basis functions for the space, even though $\mathcal{X}$ may be potentially uncountable.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

There are two settings in which we can clearly see that this holds. The first of these is when $f(x_n, \theta) = \log p(x_n|\theta)$ and we can write the likelihood as a d-dimensional exponential family. The standard exponential family form gives us the desired basis for S_0. The second setting is when $\mathcal{X}$ is in fact a finite set. If $x \in \mathcal{X}$ can only take d values, then we can easily find a basis as before.

In the exponential family case, we can gain some more intuition on the behaviour of $J(\delta)$. Consider a full rank, d-dimensional exponential family with $f_n(\theta) = \log p(x_n|\theta)$. We can write

$$(f_1 - \bar{f})(\theta) = A(\theta)^T(B(X_1) - \mathbb{E}_p(B(X_1))) - K(X_1),$$

where X_1 is the first observed data point (corresponding to the potential f_1), $B(\cdot)$ is the (full rank) sufficient statistic and $A(\theta)$ is the natural parameter. $K(X_1)$ arises from the log-base density, and is a function of X only. If there exists μ such that $R_A \coloneqq \int A(\theta)A(\theta)^T d\mu(\theta)$ has rank d, then

$$J(\delta) \xrightarrow{\delta N \to \infty} \inf_{a \in S_0; \|a\|=1} \Pr\left(\langle a, f_1 - \bar{f}\rangle_{L^2(\mu)} > 0\right) > 0.$$

We can see this as follows. Firstly, $J(\delta)$ is monotonically increasing as $\delta N \to \infty$. Hence, we can interchange the limit with the infimum to see that

$$\lim_{\delta N \to \infty} J(\delta) = \inf_{a \in S_0; \|a\|=1} \Pr\left(\langle a, f_1 - \bar{f}\rangle_{L^2(\mu)} > \lim_{\delta N \to \infty} \frac{2r_\mu}{\sqrt{N\delta}}\right),$$

and $\lim_{\delta N \to \infty} \frac{2r_\mu}{\sqrt{N\delta}} = 0$. To see that this limit is strictly positive, note that the rank condition means that $S_0 = \text{span}\,(A_1(\theta), \ldots, A_d(\theta))$. Thus, $a \in S_0 \iff a = V^T A$ for some V. Choosing μ such that $\mathbb{E}_\mu(A) = 0$,

$$\begin{aligned}\langle a, f_1 - \bar{f}\rangle_{L^2(\mu)} &= V^T \left(\int A(\theta)A(\theta)^T d\mu(\theta)\right)(B(X_1) - \mathbb{E}_p(B(X_1))) - V^T\mathbb{E}_\mu(A)K(X_1) \\ &= V^T R_A(B(X_1) - \mathbb{E}_p(B(X_1))) \qquad \text{using our choice of } \mu.\end{aligned}$$

Thus

$$\Pr\left(\langle a, f_1 - \bar{f}\rangle_{L^2(\mu)} \geq 0\right) = \Pr\left(V^T R_A(B(X_1) - \mathbb{E}_p(B(X_1))) \geq 0\right).$$

Furthermore, $\|a\| = 1 \iff V^T R_A V = 1$. Since $\mathbb{E}_p\left(V^T R_A(B(X_1) - \mathbb{E}_p(B(X_1)))\right) = 0$,

$$\begin{aligned}&\Pr\left(V^T R_A(B(X_1) - \mathbb{E}_p(B(X_1))) \geq 0\right) = 0 \\ &\iff V^T R_A(B(X_1) - \mathbb{E}_p(B(X_1))) = 0 \quad \text{almost surely.}\end{aligned}$$

However, this is impossible since both R_A and B are of full rank. Thus, $\forall V$, $\Pr(V) \coloneqq \Pr\left(V^T R_A(B(X_1) - \mathbb{E}_p(B(X_1))) \geq 0\right) > 0$. Now, if $\exists V_n$ such that $\Pr(V_n) \to 0$, the fact that we are on the compact space $\|a\| = 1$ means that there exists a subsequence $V_{n_k} \to V_*$ with $V_*^T R_A V_* = 1$. We would then have

$$V_{n_k}^T R_A(B(X_1) - \mathbb{E}_p(B(X_1))) \xrightarrow{d} V_*^T R_A(B(X_1) - \mathbb{E}_p(B(X_1))),$$

where $\xrightarrow{d}$ indicates convergence in distribution. Then at the limit, $\Pr(V_*) = 0$, which we have shown is impossible. Thus, we do indeed have that

$$\lim_{\delta N \to \infty} J(\delta) = \inf \Pr(V) > 0.$$

In this case, we can set $\delta = \omega(1/N)$, such that with probability $\geq 1 - \omega(1/N)$, we indeed have that $\min_{w \in \mathcal{W}_N} \text{KL}(\pi_w || \pi) = 0$ for $M \gtrsim d \log N$.

In order for Theorem 4.2 to hold, we require additional smoothness and concavity conditions on $f(x, \theta)$. The smoothness conditions are satisfied if, for example, $f(x; \theta)$ is C^3 in θ, as discussed in Section 4. For example, we can consider a logistic regression problem. Here, we are not in the exponential family setting. However, the logistic regression log-likelihood is concave and C^3 in θ. Hence we can find $f^{(1)}$ and $f^{(2)}$ via a second-order Taylor expansion. Finally, A4 is satisfied by any prior that is strictly positive and finite, such as the Cauchy prior we use in our logistic regression experiment.

However, the logistic log-likelihood is concave but not strictly concave. Thus A1 does not hold, and our logistic regression experiment is not covered by the assumptions of Theorem 4.2. It is therefore reassuring to see that our method obtains favourable empirical results in this experiment, and we conjecture that the strong concavity condition is sufficient but not necessary. We leave the relaxing of this condition to future work.

In the statement of Theorem 4.3, we suppose that $\exists\, \epsilon \in [0, 1)$ and $\delta \geq 0$ such that for all $w \in W$

$$\| \left(G(w)+\tau I\right)^{-1} H(w)(1-w^\star)\| \leq \epsilon \|w-w^\star\|+\delta, \tag{18}$$

where $w^\star = \text{proj}_{W^\star}(w)$ and $W^\star \subseteq \text{argmin}_{w\in W} \text{KL}(\pi_w||\pi)$.

If we can show that

$$\forall \theta \in \Theta, \quad \sum_{m=1}^{M} w_m^\star f_m(\theta) = \sum_n f_n(\theta), \tag{19}$$

then

$$\begin{aligned} H(w)(1-w^\star) &= \text{Cov}_w \left[g, f^T(1-w^\star)\right] \\ &= 0. \end{aligned} \tag{20}$$

and so thus the assumption will hold with $\epsilon = \delta = 0$.

Theorem 4.1 tells us that $\pi_{w^\star}(\theta) = \pi(\theta)$ for almost all values of θ. This implies that Eq. (19) also holds for almost all values of θ, and so Eq. (20) holds. This means that the assumption in Theorem 4.3 does in fact hold with $\epsilon = \delta = 0$.

Thus, in any setting where the conditions of Theorem 4.1 hold (such as the two discussed above), Theorem 4.3 holds as well. This reasoning also gives us some intuition into how we can generalise this. Intuitively, we require that

$$\forall \theta \in \Theta, \quad \sum_{m=1}^{M} w_m^\star f_m(\theta) \approx \sum_n f_n(\theta), \tag{21}$$

so that $H(w)(1-w^\star) \approx 0$. Essentially, we see that the better the quality of the optimal coreset $w^\star$, the tighter the bound we can find for Eq. (18). We conjecture that if $\exists\, \xi$ such that we can uniformly bound

$$\sup_{\theta\in\Theta} \left| \sum_{m=1}^{M} w_m^\star f_m(\theta) - \sum_n f_n(\theta) \right| \leq \xi,$$

then we can find ϵ and δ such that Eq. (18) holds. We leave the theoretical examination of this claim for future work.

B Proofs

In this section we give the full proofs of Theorems 4.1 to 4.3. The proofs of Theorems 4.1 and 4.2 are quite technical, so in each case we give start by giving a roadmap of the proof ideas.

B.1 Proof of Theorem 4.1

Roadmap.

1. We start the proof by giving in Eq. (22) a sufficient condition for our desired result to hold. Proving that this condition holds will be the target of the rest of the proof.
2. The next step is to show that this is indeed a sufficient condition. We do this by deriving the form for the KL divergence found in Eq. (25).
3. We then proceed by lower bounding the probability that the quantity on the right side of Eq. (22) lies within a ball of a given radius.
4. Next, we find a lower bound for the probability that we can get the left hand side of Eq. (22) to be equal to any value within the same ball as above.

5. If the events introduced in 2. and 3. hold, then we will be able to satisfy our sufficient condition Eq. (22). This implies that our desired result will also hold. Thus we can combine the two relevant lower bounds detailed in 2. and 3. to get a lower bound on the probability of our desired result holding.
6. Obtaining this final lower bound requires setting a condition on the coreset size M. We conclude the proof by rewriting condition to make it more interpretable.

Proof. We select M indices uniformly at random from $[N]$, corresponding to M potential functions $f_n(\theta)$. Because the functions f_n are generated i.i.d. and the indices chosen uniformly, the particular indices selected are unimportant for this proof; without loss of generality, we can assume the M selected indices are $n = 1, \ldots, M$.

Let μ be a probability measure such that S_0 is the associated d dimensional subspace of $L^2(\mu)$ and let $M \in [N]$. In what follows, we assume that all norms and inner products are the weighted $L^2(\mu)$ norms and inner products respectively, unless otherwise stated. Further, define $\widetilde{f}_N(\cdot) = \sum_{n=1}^{N}[f_n - \bar{f}](\cdot)/N$. We show below that with probability $\geq 1-\delta$, there exists $v_1, \cdots, v_M \geq 0$ with $\sum_m v_m = 1$ such that

$$\sum_{m=1}^{M}(f_m - \bar{f})v_m = \widetilde{f}_N. \tag{22}$$

For such a $v = (v_1, \cdots, v_M)$, set $w(v) = Nv$ (with the implicit convention that $w(v)_j = 0$ for $j > M$). Our goal is to prove that $\mathrm{KL}(\pi_{w(v)}||\pi) = 0$. We do this by starting with the form for the KL divergence derived in the proof of Lemma 3 of Campbell and Beronov [1]:

$$\mathrm{KL}(\pi_{w(v)}||\pi) = 2(1-w)^T \mathbb{E}\left[C(\gamma(T))\right](1-w),$$

where:

- $T \sim \mathrm{Beta}(1,2)$.
- $\gamma(t)$ is the path defined by $\gamma(t) = (1-t)w + t1$ for $t \in [0,1]$, coreset weights $w \in \mathbb{R}^N$, and unit vector $1 \in \mathbb{R}^N$.
- $C(w)$ is the Fisher information metric Amari [2, p. 33,34] defined as:

$$\begin{aligned} C(w) &:= \mathrm{Cov}_w[f, f] \\ &:= \mathbb{E}_{\pi_w}\left[\left(f - \mathbb{E}_{\pi_w}(f)\right)\left(f - \mathbb{E}_{\pi_w}(f)\right)^T\right]. \end{aligned}$$

  Note that we have changed the notation from that of [1], to avoid a clash of notation with terms already defined in our work.

Using the fact that the p.d.f. of a $\mathrm{Beta}(1,2)$ distribution is given by $f_T(t) = 2(1-t)$, we therefore have that

$$\begin{aligned} \mathrm{KL}(\pi_{w(v)}||\pi) &= 2(1-w)^T \mathbb{E}\left[C(\gamma(T))\right](1-w) \\ &= 2(1-w)^T \left[\int_0^1 2(1-t)C(\gamma(t))\mathrm{d}t\right](1-w) \\ &= 4(1-w)^T \left[\int_0^1 (1-t)\mathbb{E}_{w(v),t}\left[\left(f - \mathbb{E}_{\pi_{\gamma(t)}}(f)\right)\left(f - \mathbb{E}_{\pi_{\gamma(t)}}(f)\right)^T\right]\mathrm{d}t\right](1-w) \\ &= 4\int_0^1 (1-t)\mathbb{E}_{w(v),t}\left[\left((1-w)^T\left(f - \mathbb{E}_{w(v),t}(f)\right)\right)^2\right]\mathrm{d}t \\ &= 4\int_0^1 (1-t)\mathbb{E}_{w(v),t}\left[\left(\sum_{n=1}^{N}(1-w_n)f_n - \mathbb{E}_{w(v),t}\left(\sum_{n=1}^{N}(1-w_n)f_n\right)\right)^2\right]\mathrm{d}t, \end{aligned} \tag{23}$$

where $\mathbb{E}_{w(v),t}$ refers to taking expectations under $\pi_{w(v),t}$, the coreset posterior corresponding to weights given by $\gamma(t) = (1-t)w + t1$. To be explicit:

$$\pi_{w(v),t}(\theta) = \frac{e^{t\sum_n f_n + (1-t)\sum_n w_n f_n}\pi_0(\theta)}{\int_\Theta e^{t\sum_n f_n + (1-t)\sum_n w_n f_n}\pi_0(\theta)d\theta}. \tag{24}$$

We now note that

$$\begin{aligned}\frac{1}{N}\sum_{n=1}^N (1-w_n)f_n &= \frac{1}{N}\sum_{n=1}^N (f_n - \bar{f}) - \frac{1}{N}\sum_{n=1}^N w_n(f_n - \bar{f}) \\ &= \widetilde{f}_N - \sum_m v_m(f_m - \bar{f}),\end{aligned}$$

and thus

$$\begin{aligned}t\sum_n f_n + (1-t)\sum_n w_n f_n &= \sum_n f_n - (1-t)\sum_n (1-w_n)f_n \\ &= \sum_n f_n - N(1-t)[\widetilde{f}_N - \sum_m v_m(f_m - \bar{f})].\end{aligned}$$

Thus, by substituting into Eq. (24), we can write $\mathbb{E}_{w(v),t}$ as

$$\mathbb{E}_{w(v),t}(h) = \frac{\int_\Theta h(\theta)e^{\sum_n f_n(\theta) - N(1-t)[\widetilde{f}_N - \sum_m v_m(f_m - \bar{f})]}\pi_0(\theta)d\theta}{\int_\Theta e^{\sum_n f_n(\theta) - N(1-t)[\widetilde{f}_N - \sum_m v_m(f_m - \bar{f})]}\pi_0(\theta)d\theta}.$$

Furthermore, substituting into Eq. (23) we have that

$$\mathrm{KL}(\pi_{w(v)}||\pi) = 4N^2\int_0^1 (1-t)\mathbb{E}_{w(v),t}\left[\left(\widetilde{f}_N - \sum_m v_m(f_m - \bar{f}) - \mathbb{E}_{w(v),t}\left(\widetilde{f}_N - \sum_m v_m(f_m - \bar{f})\right)\right)^2\right]\mathrm{d}t. \tag{25}$$

If we can prove Eq. (22), Eq. (25) would give us that $\mathrm{KL}(\pi_{w(v)}||\pi) = 0$, thus completing the proof. Of course, if Eq. (22) holds then the coreset and full posteriors are equal, so $\mathrm{KL}(\pi_{w(v)}||\pi)$ is trivially zero. However, targeting this expression will also be helpful for the proof of Theorem 4.2. In order to prove Eq. (22), first note that $\|\widetilde{f}_N\|^2_{L^2(\mu)} = \mathbb{E}_\mu\left(\left(\frac{1}{N}\sum_n (f_n - \bar{f})\right)^2\right)$. With $r_\mu^2 = \mathbb{E}_\mu\left(\mathbb{E}_p([f_1 - \bar{f}]^2)\right)$, we have by Markov's inequality that, $\forall \delta > 0$,

$$\begin{aligned}\Pr\left(\|\widetilde{f}_N\|^2_{L^2(\mu)} > \frac{r_\mu^2}{N\delta}\right) &\leq \frac{N\delta}{r_\mu^2}\mathbb{E}_p\left(\mathbb{E}_\mu\left(\left(\frac{1}{N}\sum_n (f_n - \bar{f})\right)^2\right)\right) \\ &= \frac{N\delta}{r_\mu^2}\mathbb{E}_\mu\left(\mathbb{E}_p\left(\left(\frac{1}{N}\sum_n (f_n - \bar{f})\right)^2\right)\right) \\ &= \frac{N\delta}{r_\mu^2}\frac{1}{N}\mathbb{E}_\mu\left(\mathbb{E}_p\left(\left(f_1 - \bar{f}\right)^2\right)\right) \\ &= \delta.\end{aligned}$$

Defining $t_N \coloneqq \sqrt{r_\mu^2/(N\delta)}$, it is enough to show that, with high probability, the convex hull of $(f_m - \bar{f}), m \in [M]$ contains the ball (in S_0) centered at 0 and with radius t_N. As in the theorem statement, S_0 is defined as the vector space of functions on Θ spanned by $\{f(x,\cdot); x \in \mathcal{X}\}$. This boils down to bounding

$$\Pr\left(\inf_{a \in S_0; \|a\|=1}\max_m \langle a, f_m - \bar{f}\rangle \geq t_N\right)$$

from below, since on this event any point in the ball will be in the convex hull. Moreover, with probability $\geq 1-\delta$, $\widetilde{f}_N$ is in this ball, and thus Eq. (22) will be satisfied.

From Böröczky and Wintsche [3, Corollary 1.2], the unit sphere in d-dimensions can be covered by

$$N_d(\phi) = \frac{C \cdot \cos\phi}{\sin^d \phi} d^{\frac{3}{2}} \log(1 + d\cos^2\phi) \leq \phi^{-d} A_d, \quad A_d = C e^{\frac{d}{2}} d^{\frac{3}{2}} \log(1+d)$$

balls of radius $0 < \phi \leq \arccos \frac{1}{\sqrt{d+1}}$ and centers $(a_i)_{i \leq N_d(\phi)}$, where C is a universal constant. Moreover,

$$\inf_{a \in S_0; \|a\|=1} \max_m \langle a, f_m - \bar{f} \rangle \geq \min_{i \leq N_d(\phi)} \max_{m=1,\ldots,M} \left[\langle a_i, f_m - \bar{f} \rangle - \phi \|f_m - \bar{f}\|_{L^2(\mu)} \right].$$

For any $a \in S_0$ with $\|a\| = 1$, by independence,

$$\begin{aligned} &\Pr\left(\max_{m=1,\ldots,M} \left[\langle a, f_m - \bar{f} \rangle - \phi \|f_m - \bar{f}\|_{L^2(\mu)} \right] \leq t_N \right) \\ &\quad = \Pr\left(\langle a, f_1 - \bar{f} \rangle - \phi \|f_1 - \bar{f}\|_{L^2(\mu)} \leq t_N \right)^M. \end{aligned}$$

For any $\phi > 0$, if $\langle a, f_1 - \bar{f} \rangle \leq 2t_N$ and $\|f_1 - \bar{f}\|_{L^2(\mu)} > t_N/\phi$ then necessarily $\langle a, f_1 - \bar{f} \rangle - \phi \|f_1 - \bar{f}\|_{L^2(\mu)} \leq t_N$ by the triangle inequality. Thus, we can use the union bound to bound the above by

$$\Pr\left(\langle a, f_1 - \bar{f} \rangle - \phi \|f_1 - \bar{f}\|_{L^2(\mu)} \leq t_N \right)^M \leq \left[\Pr\left(\langle a, f_1 - \bar{f} \rangle \leq 2t_N \right) + \Pr\left(\|f_1 - \bar{f}\| > t_N/\phi \right) \right]^M$$

We have, for all $a \in S_0$ with $\|a\| = 1$,

$$\begin{aligned} \Pr(\langle a, f_1 - \bar{f} \rangle > 2t_N) &\geq \inf_{a \in S_0; \|a\|=1} \Pr\left(\langle a, f_1 - \bar{f} \rangle > 2t_N \right) \\ &= J(\delta), \end{aligned}$$

using the definition of $J(\delta)$ as given in the statement of Theorem 4.1. Thus,

$$\Pr(\langle a, f_1 - \bar{f} \rangle \leq 2t_N) \leq 1 - J(\delta).$$

Furthermore, choosing $\phi^2 = J(\delta)/(4N\delta)$, we have by Chebychev's inequality that

$$\Pr\left(\|f_1 - \bar{f}\| > t_N/\phi \right) \leq J(\delta)/2.$$

Finally we obtain, using the union bound and the above results, that

$$\begin{aligned} \Pr\left(\inf_{a \in S_0; \|a\|=1} \max_m \langle a, f_m - \bar{f} \rangle < t_N \right) &\leq \Pr\left(\min_{i \leq N_d(\phi)} \max_{m=1,\ldots,M} \left[\langle a_i, f_m - \bar{f} \rangle - \phi \|f_m - \bar{f}\|_{L^2(\mu)} \right] < t_N \right) \\ &\leq \sum_{i \leq N_d(\phi)} \Pr\left(\max_{m=1,\ldots,M} \left[\langle a_i, f_m - \bar{f} \rangle - \phi \|f_m - \bar{f}\|_{L^2(\mu)} \right] < t_N \right) \\ &\leq \sum_{i \leq N_d(\phi)} \left[\Pr\left(\langle a_i, f_1 - \bar{f} \rangle \leq 2t_N \right) + \Pr\left(\|f_1 - \bar{f}\| > t_N/\phi \right) \right]^M \\ &\leq \sum_{i \leq N_d(\phi)} (1 - J(\delta) + J(\delta)/2)^M \\ &= N_d(\phi)(1 - J(\delta)/2)^M. \end{aligned}$$

For M such that $MJ(\delta) \geq 4 \log\left(N_d(\phi)\right)$, we then have that

$$\begin{aligned} N_d(\phi)(1 - J(\delta)/2)^M &\leq e^{MJ(\delta)/4} e^{M \log(1 - J(\delta)/2)} \\ &\leq e^{MJ(\delta)/4} e^{-MJ(\delta)/2} = e^{-MJ(\delta)/4}. \end{aligned}$$

Using the union bound, we can therefore see that Eq. (22) holds with probability $\geq 1 - \delta - e^{-MJ(\delta)/4}$.

Noting that $N_d(\phi) \leq \phi^{-d} A_d$, a sufficient condition on M for this to hold is that

$$\begin{aligned} M &\geq \frac{4}{J(\delta)} \log\left(\phi^{-d} A_d \right) \\ &= \frac{4}{J(\delta)} \log\left(\left(\frac{J(\delta)}{4N\delta} \right)^{-d/2} A_d \right) \\ &= \frac{2d}{J(\delta)} \left[\log N + \log\left(\frac{4\delta}{J(\delta)} \right) + \log\left(A_d^{2/d} \right) \right] \\ &\geq \frac{2d}{J(\delta)} \left[\log N + \log\left(\frac{4\delta}{J(\delta)} \right) + C_1 \right], \end{aligned}$$

where C_1 is a constant such that $C_1 \leq \log\left(A_d^{2/d}\right)$. This arrives at the condition on M in the theorem statement, and thus completes the proof. □

B.2 Proof of Theorem 4.2

Roadmap.

1. The goal of this proof is to upper bound Eq. (25). We start by using a substitution to find an initial upper bound in the form of an integral.
2. Next, we split this integral up into two different terms, which we call $I_{\epsilon,1}$ and $I_{\epsilon,2}$, that together sum to the full integral as shown in Eq. (26). These terms can both be expressed as fractions, and share a common denominator. We will then find upper bound for these two terms separately.
3. We start by targeting $I_{\epsilon,1}$, and in particular upper bounding its numerator. We do this across a series of steps which provide sequential upper bounds. The aim is to find a final upper bound which is simpler, while retaining the necessary asymptotic properties. The final upper bound is given in Eq. (31).
4. As part of this process, we assume that a certain set of events occur, and lower bound the probability that they do. These probability bounds will be incorporated into the high probability bound for the final result.
5. Obtaining this upper bound on the numerator of $I_{\epsilon,1}$ also requires the introduction a condition similar to the one that was central to the proof of Theorem 4.1. This condition is given in Eq. (27).
6. Next, we lower bound the common denominator of $I_{\epsilon,1}$ and $I_{\epsilon,2}$. This involves a series of steps similar to those involved in upper bounding the numerator of $I_{\epsilon,1}$, and again we need to assume a set of events occurring. As before, we lower bound the probabilities that they do occur, and these probability bounds will be incorporated into the high probability bound for the final result. The final lower bound is given in Eq. (40).
7. Similarly, we need to introduce two more important conditions that are required for our lower bound on the denominator to hold. These are given BY Eqs. (32) and (35).
8. So far, we have obtained in Eq. (44) an upper bound on $I_{\epsilon,1}$, which holds with a probability that we have obtained a lower bound for, under certain specific conditions that we specify in Eqs. (41) to (43).
9. The next step is to upper bound $I_{\epsilon,2}$, and in particular its numerator (since we have already lower bounded its denominator). We do this using similar techniques to before, which introduces additional terms into the final high probability bounds. The final upper bound is given in Eq. (45).
10. This gives us an upper bound on our original target (i.e. Eq. (25)), which holds with a probability that we have lower bounded, and assuming that the conditions Eqs. (41) to (43) hold.
11. The final step is to lower bound the probability that Eqs. (41) to (43) all hold. This can be done in the same way as in the proof for Theorem 4.1, and we omit the full details.
12. Incorporating the above probability lower bounds, we finally have our desired upper bound for the original target, and a lower bound for the probability that this bound holds. This is what we refer to as our "high probability bound" in this roadmap.

Proof. By the concavity and differentiability of $f(x,\cdot)$ we have $f(x;\theta) \leq f(x;\theta_0) + \nabla_\theta f(x;\theta_0)^T(\theta-\theta_0) - L_0\|\theta-\theta_0\|^2$. Set $S_N(\theta_0) = \sum_n \nabla_\theta f_n(\theta_0)/\sqrt{N}$ and recall that by definition of θ_0 as the maximizer of $\bar{f} = \mathbb{E}_p(f_1(\theta))$, $\mathbb{E}_p\left(\nabla_\theta f_1(\theta_0)\right) = 0$.

Starting from Eq. (25), we first use the fact that $\widetilde{f}_N(\cdot) = \sum_{n=1}^N [f_n - \bar{f}](\cdot)/N$ to rewrite

$$\widetilde{f}_N - \sum_m v_m(f_m - \bar{f}) - \mathbb{E}_{w(v),t}[\widetilde{f}_N - \sum_m v_m(f_m - \bar{f})]$$

$$= \frac{\sum_n f_n(\theta) - f_n(\theta_0)}{N} - \sum_m v_m(f_m(\theta) - f_m(\theta_0)) - \mathbb{E}_{w(v),t}\left[\frac{\sum_n f_n(\theta) - f_n(\theta_0)}{N} - \sum_m v_m(f_m(\theta) - f_m(\theta_0))\right].$$

Substituting this into Eq. (25), and using the fact that, for any random variable X, $\mathbb{E}\left[(X - \mathbb{E}(X))^2\right] = \mathbb{E}\left(X^2\right) - \mathbb{E}\left(X\right)^2 \leq \mathbb{E}\left(X^2\right)$, we can upper bound Eq. (25) by

$$\mathrm{KL}(\pi_{w(v)}||\pi) \leq 4N^2 \int_0^1 (1-t)\mathbb{E}_{w(v),t}\left[\left(\frac{1}{N}\sum_n f_n(\theta) - f_n(\theta_0) - \sum_m v_m(f_m(\theta) - f_m(\theta_0))\right)^2\right] \mathrm{d}t.$$

We now decompose, for each $t \in (0,1)$, the integral over Θ in the above into an integral over $B_\epsilon = \{\|\theta - \theta_0\| \leq \epsilon\}$ and an integral over B_ϵ^c. Here, $0 < \epsilon < \epsilon_0$, where ϵ_0 is as defined in the assumptions of Theorem 4.2. We can rewrite $\mathbb{E}_{w(v),t}(h)$ as

$$\begin{aligned}
\mathbb{E}_{w(v),t}(h) &= \frac{\int_\Theta h(\theta) e^{\sum_n f_n(\theta) - N(1-t)[\widetilde{f}_N(\theta) - \sum_m v_m(f_m(\theta) - \bar{f}(\theta))]}\pi_0(\theta)d\theta}{\int_\Theta e^{\sum_n f_n(\theta) - N(1-t)[\widetilde{f}_N(\theta) - \sum_m v_m(f_m(\theta) - \bar{f}(\theta))]}\pi_0(\theta)d\theta} \\
&= \frac{\int_\Theta h(\theta) e^{t\sum_n f_n(\theta) + N(1-t)\sum_m v_m f_m(\theta)}\pi_0(\theta)d\theta}{\int_\Theta e^{t\sum_n f_n(\theta) + N(1-t)\sum_m v_m f_m(\theta)}\pi_0(\theta)d\theta} \times \frac{e^{-t\sum_n f_n(\theta_0) - N(1-t)\sum_m v_m f_m(\theta_0)}}{e^{-t\sum_n f_n(\theta_0) - N(1-t)\sum_m v_m f_m(\theta_0)}} \\
&= \frac{\int_\Theta h(\theta) e^{t\sum_n (f_n(\theta) - f_n(\theta_0)) + N(1-t)\sum_m v_m(f_m(\theta) - f_m(\theta_0))]}\pi_0(\theta)d\theta}{\int_\Theta e^{t\sum_n (f_n(\theta) - f_n(\theta_0)) + N(1-t)\sum_m v_m(f_m(\theta) - f_m(\theta_0))]}\pi_0(\theta)d\theta} \\
&= \frac{\int_{B_\epsilon^c} h(\theta) e^{t\sum_n (f_n(\theta) - f_n(\theta_0)) + N(1-t)\sum_m v_m(f_m(\theta) - f_m(\theta_0))]}\pi_0(\theta)d\theta}{\int_\Theta e^{t\sum_n (f_n(\theta) - f_n(\theta_0)) + N(1-t)\sum_m v_m(f_m(\theta) - f_m(\theta_0))]}\pi_0(\theta)d\theta} \\
&\quad + \frac{\int_{B_\epsilon} h(\theta) e^{t\sum_n (f_n(\theta) - f_n(\theta_0)) + N(1-t)\sum_m v_m(f_m(\theta) - f_m(\theta_0))]}\pi_0(\theta)d\theta}{\int_\Theta e^{t\sum_n (f_n(\theta) - f_n(\theta_0)) + N(1-t)\sum_m v_m(f_m(\theta) - f_m(\theta_0))]}\pi_0(\theta)d\theta} \\
&\coloneqq I_{\epsilon,1}(h) + I_{\epsilon,2}(h)
\end{aligned}$$

Writing $I_{\epsilon,1} = I_{\epsilon,1}\left(\left(\frac{1}{N}\sum_n f_n(\theta) - f_n(\theta_0) - \sum_m v_m(f_m(\theta) - f_m(\theta_0))\right)^2\right)$, and similarly for $I_{\epsilon,2}$, we can thus upper bound Eq. (25) by

$$KL(\pi_{w(v)}||\pi) \leq 4N^2 \int_0^1 (1-t)\left(I_{\epsilon,1} + I_{\epsilon,2}\right) \mathrm{d}t \tag{26}$$

We first prove that the integral $I_{\epsilon,1}$ over B_ϵ^c is small. Using the concavity bound for f_n, and recalling that $\sum_m v_m = 1$, we have that

$$\begin{aligned}
&t\sum_n (f_n(\theta) - f_n(\theta_0)) + N(1-t)\sum_m v_m(f_m(\theta) - f_m(\theta_0)) \\
&\qquad \leq t\sum_n \left(\nabla_\theta f_n(\theta_0)^T(\theta - \theta_0) - L_0\|\theta - \theta_0\|^2\right) + N(1-t)\sum_m v_m\left(\nabla_\theta f_m(\theta_0)^T(\theta - \theta_0) - L_0\|\theta - \theta_0\|^2\right) \\
&\qquad = t\left[\sqrt{N}S_N(\theta_0)^T(\theta - \theta_0) - NL_0\|\theta - \theta_0\|^2\right] + N(1-t)\left[\sum_m v_m \nabla_\theta f_m(\theta_0)^T(\theta - \theta_0) - L_0\|\theta - \theta_0\|^2\right].
\end{aligned}$$

If

$$\sum_m v_m \nabla_\theta f_m(\theta_0) = \frac{S_N(\theta_0)}{\sqrt{N}} \tag{27}$$

then

$$t\sum_n (f_n(\theta) - f_n(\theta_0)) + N(1-t)\sum_m v_m(f_m(\theta) - f_m(\theta_0)) \leq \sqrt{N}S_N(\theta_0)^T(\theta - \theta_0) - NL_0\|\theta - \theta_0\|^2.$$

Using this bound, we can bound $I_{\epsilon,1}$ by

$$
\begin{aligned}
&I_{\epsilon,1} \\
&:= \frac{\int_{B_\epsilon^c} \left[\frac{1}{N}\sum_n f_n(\theta) - f_n(\theta_0) - \sum_m v_m(f_m(\theta) - f_m(\theta_0))\right]^2 e^{t\sum_n (f_n(\theta)-f_n(\theta_0)) + N(1-t)\sum_m v_m(f_m(\theta)-f_m(\theta_0))]}\pi_0(\theta)d\theta}{\int_\Theta e^{t\sum_n (f_n(\theta)-f_n(\theta_0)) + N(1-t)\sum_m v_m(f_m(\theta)-f_m(\theta_0))]}\pi_0(\theta)d\theta} \\
&\leq \frac{\int_{B_\epsilon^c} \left[\frac{1}{N}\sum_n f_n(\theta) - f_n(\theta_0) - \sum_m v_m(f_m(\theta) - f_m(\theta_0))\right]^2 e^{\sqrt{N}S_N(\theta_0)^T(\theta-\theta_0) - NL_0\|\theta-\theta_0\|^2}\pi_0(\theta)d\theta}{\int_\Theta e^{t\sum_n (f_n(\theta)-f_n(\theta_0)) + N(1-t)\sum_m v_m(f_m(\theta)-f_m(\theta_0))]}\pi_0(\theta)d\theta}.
\end{aligned}
$$

Completing the square in the exponent of the numerator, we can rewrite it as

$$
\begin{aligned}
&\sqrt{N}S_N(\theta_0)^T(\theta-\theta_0) - NL_0\|\theta-\theta_0\|^2 \\
&= -NL_0\left(\|\theta-\theta_0\|^2 - \frac{S_N(\theta_0)^T(\theta-\theta_0)}{\sqrt{N}L_0}\right) \\
&= -NL_0\left\|(\theta-\theta_0) - \frac{S_N(\theta_0)}{2\sqrt{N}L_0}\right\|^2 + \frac{\|S_N(\theta_0)\|^2}{4L_0}
\end{aligned}
$$

We now assume that $\|S_N(\theta_0)\| \leq \left(2-\sqrt{2}\right)\epsilon L_0\sqrt{N}$. By Markov's inequality,

$$
\begin{aligned}
\Pr\left(\|S_N(\theta_0)\| \geq \left(2-\sqrt{2}\right)\epsilon L_0\sqrt{N}\right) &\leq \frac{\mathbb{E}_p\left(\|S_N(\theta_0)\|^2\right)}{\left(2-\sqrt{2}\right)^2\epsilon^2 N L_0^2} \\
&\leq 3\frac{\mathbb{E}_p\left(\|\sum_n \nabla_\theta f_n(\theta_0)\|^2\right)}{\epsilon^2 N^2 L_0^2} \\
&= 3\frac{\sum_n \mathbb{E}_p\left(\|\nabla_\theta f_n(\theta_0)\|^2\right)}{\epsilon^2 N^2 L_0^2} \\
&= 3\frac{\mathbb{E}_p\left(\|\nabla_\theta f_1(\theta_0)\|^2\right)}{\epsilon^2 N L_0^2},
\end{aligned}
$$

and thus the probability that $\|S_N(\theta_0)\| \leq \left(2-\sqrt{2}\right)\epsilon L_0\sqrt{N}$ is bounded from below by $1 - 3\mathbb{E}_p(\|\nabla_\theta f_1(\theta_0)\|^2)/(L_0^2 N\epsilon^2)$. Here, and throughout, we assume that N is large enough such that the relevant probabilities we use for our high-probability bounds are ≥ 0. When this event occurs,

$$
\begin{aligned}
\left\|(\theta-\theta_0) - \frac{S_N(\theta_0)}{2\sqrt{N}L_0}\right\| &\geq \|\theta-\theta_0\| - \left\|\frac{S_N(\theta_0)}{2\sqrt{N}L_0}\right\| \\
&\geq \|\theta-\theta_0\| - \frac{\left(2-\sqrt{2}\right)\epsilon L_0\sqrt{N}}{2\sqrt{N}L_0} \\
&\geq \|\theta-\theta_0\| - \left(\frac{2-\sqrt{2}}{2}\right)\epsilon \\
&\geq \|\theta-\theta_0\| - \left(\frac{2-\sqrt{2}}{2}\right)\|\theta-\theta_0\| \qquad \text{on } B_\epsilon^c \\
&= \frac{1}{\sqrt{2}}\|\theta-\theta_0\|.
\end{aligned}
$$

Thus

$$
\left\|(\theta-\theta_0) - \frac{S_N(\theta_0)}{2\sqrt{N}L_0}\right\|^2 \geq \frac{1}{2}\|\theta-\theta_0\|^2,
$$

since both sides are positive numbers.

For ease of notation, we define $\Delta_n \coloneqq f_n(\theta) - f_n(\theta_0)$. We can then bound $I_{\epsilon,1}$ by

$$\begin{aligned} I_{\epsilon,1} &\leq \frac{e^{\frac{\|S_N(\theta_0)\|^2}{4L_0}} \int_{B_\epsilon^c} \left[\frac{1}{N}\sum_n \Delta_n - \sum_m v_m \Delta_m\right]^2 e^{-NL_0 \left\|(\theta-\theta_0) - \frac{S_N(\theta_0)}{2\sqrt{N}L_0}\right\|^2} \pi_0(\theta) d\theta}{\int_\Theta e^{t\sum_n (f_n(\theta)-f_n(\theta_0)) + N(1-t)\sum_m v_m (f_m(\theta)-f_m(\theta_0))] \pi_0(\theta) d\theta} \\ &\leq \frac{e^{\epsilon^2 NL_0/8} \int_{B_\epsilon^c} \left[\frac{1}{N}\sum_n \Delta_n - \sum_m v_m \Delta_m\right]^2 e^{-NL_0\|\theta-\theta_0\|^2/2} \pi_0(\theta) d\theta}{\int_\Theta e^{t\sum_n (f_n(\theta)-f_n(\theta_0)) + N(1-t)\sum_m v_m (f_m(\theta)-f_m(\theta_0))] \pi_0(\theta) d\theta}. \end{aligned}$$

To further simplify the upper bound on the numerator, we define $\bar{\Delta} \coloneqq \mathbb{E}_p(\Delta_n)$ and consider the event where

$$\int_{B_\epsilon^c} \left[\frac{1}{N}\sum_n \Delta_n - \sum_m v_m \Delta_m\right]^2 e^{-NL_0\|\theta-\theta_0\|^2/2} \pi_0(\theta) d\theta \leq e^{-NL_0\epsilon^2/4} \int_\Theta \mathbb{E}_p(\Delta_1 - \bar{\Delta})^2(\theta) \pi_0(\theta) d\theta. \tag{28}$$

The probability of Eq. (28) holding can be bounded from below by using Markov's inequality,

$$\begin{aligned} &\Pr\left(\int_{B_\epsilon^c} \left[\frac{1}{N}\sum_n \Delta_n - \sum_m v_m \Delta_m\right]^2 e^{-NL_0\|\theta-\theta_0\|^2/2} \pi_0(\theta) d\theta > e^{-NL_0\epsilon^2/4} \int_\Theta \mathbb{E}_p(\Delta_1 - \bar{\Delta})^2(\theta) \pi_0(\theta) d\theta\right) \\ &\leq \frac{e^{NL_0\epsilon^2/4} \mathbb{E}_p\left(\int_{B_\epsilon^c} \left[\frac{1}{N}\sum_n \Delta_n - \sum_m v_m \Delta_m\right]^2 e^{-NL_0\|\theta-\theta_0\|^2/2} \pi_0(\theta) d\theta\right)}{\int_\Theta \mathbb{E}_p(\Delta_1 - \bar{\Delta})^2(\theta) \pi_0(\theta) d\theta}. \end{aligned} \tag{29}$$

We now use the fact that $(a-b)^2 \leq 2a^2 + 2b^2$ to bound

$$\begin{aligned} \left[\frac{1}{N}\sum_n \Delta_n - \sum_m v_m \Delta_m\right]^2 &= \left[\frac{\sum_n (\Delta_n - \bar{\Delta})}{N} - \sum_m v_m (\Delta_m - \bar{\Delta})\right]^2 \\ &\leq 2\left(\frac{\sum_n (\Delta_n - \bar{\Delta})}{N}\right)^2 + 2\left(\sum_m v_m (\Delta_m - \bar{\Delta})\right)^2 \\ &\leq 2\left(\frac{\sum_n (\Delta_n - \bar{\Delta})}{N}\right)^2 + 2\left(\max_M (\Delta_m - \bar{\Delta}) \sum_m v_m\right)^2 \\ &\leq 2\left(\frac{\sum_n (\Delta_n - \bar{\Delta})}{N}\right)^2 + 2\max_M (\Delta_m - \bar{\Delta})^2. \end{aligned}$$

Combined with the fact that $\|\theta - \theta_0\| \geq \epsilon$ on B_ϵ^c, we can bound Eq. (29) above by

$$\leq \frac{2\left(\int_{B_\epsilon^c} \mathbb{E}_p\left[\left(\frac{\sum_n (\Delta_n - \bar{\Delta})}{N}\right)^2 + \max_M (\Delta_m - \bar{\Delta})^2\right] e^{-NL_0\|\theta-\theta_0\|^2/4} \pi_0(\theta) d\theta\right)}{\int_\Theta \mathbb{E}_p(\Delta_1 - \bar{\Delta})^2(\theta) \pi_0(\theta) d\theta}. \tag{30}$$

Then, by the definition of $\bar{\Delta}$, and the independence of the f_n and therefore Δ_n,

$$\begin{aligned} \mathbb{E}_p\left[\left(\frac{\sum_n (\Delta_n - \bar{\Delta})}{N}\right)^2 + \max_M (\Delta_m - \bar{\Delta})^2\right] &= \frac{1}{N^2}\sum_n \mathbb{E}_p\left((\Delta_n - \bar{\Delta})^2\right) + \mathbb{E}_p\left[\max_M (\Delta_m - \bar{\Delta})^2\right] \\ &\leq \frac{1}{N}\mathbb{E}_p\left((\Delta_1 - \bar{\Delta})^2\right) + \mathbb{E}_p\left[\sum_{m=1}^M (\Delta_m - \bar{\Delta})^2\right] \\ &= \frac{1}{N}\mathbb{E}_p\left((\Delta_1 - \bar{\Delta})^2\right) + M\mathbb{E}_p\left((\Delta_1 - \bar{\Delta})^2\right). \end{aligned}$$

Thus, Eq. (30) can be upper bounded by

$$\begin{aligned}
&\leq \frac{2\int_{B_\epsilon^c} e^{-NL_0\|\theta-\theta_0\|^2/4}\mathbb{E}_p(\Delta_1-\bar{\Delta})^2(\theta)\left(\frac{1}{N}+M\right)\pi_0(\theta)d\theta}{\int_\Theta \mathbb{E}_p(\Delta_1-\bar{\Delta})^2(\theta)\pi_0(\theta)d\theta} \\
&\leq \frac{2\int_{B_\epsilon^c} e^{-NL_0\epsilon^2/4}\mathbb{E}_p(\Delta_1-\bar{\Delta})^2(\theta)\left(2M\right)\pi_0(\theta)d\theta}{\int_\Theta \mathbb{E}_p(\Delta_1-\bar{\Delta})^2(\theta)\pi_0(\theta)d\theta} \\
&\leq 4Me^{-NL_0\epsilon^2/4}.
\end{aligned}$$

This tells us that the probability of Eq. (28) holding can be bounded from below by $1-4Me^{-NL_0\epsilon^2/4}$. We have now obtained an upper bound on the numerator of $I_{\epsilon,1}$ that holds with high probability. To summarize the result so far: if $\sum_m v_m \nabla_\theta f_m(\theta_0) = \frac{S_N(\theta_0)}{\sqrt{N}}$, we have that for any $0<\epsilon<\epsilon_0$,

$$I_{\epsilon,1} \leq \frac{e^{-\epsilon^2 NL_0/8}\int_\Theta \mathbb{E}_p(\Delta_1-\bar{\Delta})^2(\theta)\pi_0(\theta)d\theta}{\int_\Theta e^{t\sum_n(f_n(\theta)-f_n(\theta_0))+N(1-t)\sum_m v_m(f_m(\theta)-f_m(\theta_0))}\pi_0(\theta)d\theta}. \tag{31}$$

with probability at least

$$1-4Me^{-NL_0\epsilon^2/4}-3\frac{\mathbb{E}_p\left(\|\nabla_\theta f_1(\theta_0)\|^2\right)}{\epsilon^2 NL_0^2}.$$

We continue by finding a lower bound for the denominator of $I_{\epsilon,1}$. We decompose f_n into $f_n^{(1)}$ and $f_n^{(2)}$ and assume that

$$\sum_m v_m(f_m^{(1)}(\theta)-f_m^{(1)}(\theta_0)) - \frac{\sum_n(f_n^{(1)}(\theta)-f_n^{(1)}(\theta_0))}{N} = 0. \tag{32}$$

Then

$$\begin{aligned}
&t\sum_n(f_n(\theta)-f_n(\theta_0)) + N(1-t)\left[\sum_m v_m(f_m(\theta)-f_m(\theta_0))\right] \\
&= \sum_n(f_n(\theta)-f_n(\theta_0)) + N(1-t)\left[\sum_m v_m(f_m(\theta)-f_m(\theta_0)) - \frac{\sum_n(f_n(\theta)-f_n(\theta_0))}{N}\right] \\
&= \sum_n(f_n(\theta)-f_n(\theta_0)) + N(1-t)\left[\sum_m v_m(f_m^{(2)}(\theta)-f_m^{(2)}(\theta_0)) - \frac{\sum_n(f_n^{(2)}(\theta)-f_n^{(2)}(\theta_0))}{N}\right] \\
&= \sum_n(f_n(\theta)-f_n(\theta_0)) + N(1-t)\left[\sum_m v_m f_m^{(2)}(\theta) - \frac{\sum_n f_n^{(2)}(\theta)}{N}\right],
\end{aligned} \tag{33}$$

using Eq. (32) and the fact that $f_n^{(2)}(\theta_0)=0$ by assumption **A2**. For any $0<\epsilon'<\epsilon_0$, and $\|\theta-\theta_0\|\leq\epsilon'$, we use assumption **A2** again, along with the fact that $a-b\geq -|a|-|b|$, to bound Eq. (33) from below by

$$\begin{aligned}
&\geq \sum_n(f_n(\theta)-f_n(\theta_0)) - N(1-t)\left|\sum_m v_m f_m^{(2)}(\theta) - \frac{\sum_n f_n^{(2)}(\theta)}{N}\right| \\
&\geq \sum_n(f_n(\theta)-f_n(\theta_0)) - N(1-t)r(\epsilon')\|\theta-\theta_0\|^2\left[\left|\sum_m v_m R(x_m)\right| + \left|\frac{\sum_n R(x_n)}{N}\right|\right] \\
&\geq \sum_n(f_n(\theta)-f_n(\theta_0)) \\
&\quad - N(1-t)r(\epsilon')\|\theta-\theta_0\|^2\left[2\mathbb{E}_p(R(X)) + \left|\sum_m v_m(R(x_m)-\mathbb{E}_p(R(X)))\right| + \frac{|\sum_n(R(x_n)-\mathbb{E}_p(R(X)))|}{N}\right]
\end{aligned} \tag{34}$$

using the fact that $R(x) > 0$ by assumption. Next, we assume that

$$\left|\sum_m v_m(R(x_m) - \mathbb{E}_p(R(X))\right| \leq 1 \qquad \text{and} \qquad \frac{|\sum_n (R(x_n) - \mathbb{E}_p(R(X))|}{N} \leq 1. \tag{35}$$

We can bound the probability of the second inequality holding from below by $1 - \mathrm{Var}_p(R(X))/N$ via Chebychev's inequality:

$$\Pr\left(\left|\frac{1}{N}\sum_n R(x_n) - \mathbb{E}_p(R(X))\right| \geq \frac{\sqrt{\mathrm{Var}_p\left(\frac{1}{N}\sum_n R(x_n)\right)}}{\sqrt{\mathrm{Var}_p\left(\frac{1}{N}\sum_n R(x_n)\right)}}\right) \leq \mathrm{Var}_p\left(\frac{1}{N}\sum_n R(x_n)\right) = \frac{1}{N}\mathrm{Var}_p(R(X)).$$

Substituting the two assumptions from Eq. (35) into Eq. (34) above, we obtain that

$$\begin{aligned}
&t\sum_n (f_n(\theta) - f_n(\theta_0)) + N(1-t)\sum_m v_m(f_m(\theta) - f_m(\theta_0)) \\
&\qquad \geq \sum_n (f_n(\theta) - f_n(\theta_0)) - 2Nr(\epsilon')\|\theta - \theta_0\|^2 \left[\mathbb{E}_p(R(X)) + 1\right] \\
&\qquad = \sum_n (f_n(\theta) - f_n(\theta_0)) - \gamma N r(\epsilon')\|\theta - \theta_0\|^2, \quad \text{where} \quad \gamma \coloneqq 2\left[\mathbb{E}_p(R(X)) + 1\right].
\end{aligned}$$

Then we can bound the denominator of $I_{\epsilon,1}$ from below by

$$\begin{aligned}
&\int_\Theta e^{t\sum_n (f_n(\theta)-f_n(\theta_0)) + N(1-t)\sum_m v_m(f_m(\theta)-f_m(\theta_0))]}\pi_0(\theta)d\theta \\
\geq &\int_{B_{\epsilon'}} e^{t\sum_n (f_n(\theta)-f_n(\theta_0)) + N(1-t)\sum_m v_m(f_m(\theta)-f_m(\theta_0))]}\pi_0(\theta)d\theta \\
\geq &\int_{B_{\epsilon'}} e^{\sum_n f_n(\theta)-f_n(\theta_0) - \gamma N r(\epsilon')\|\theta-\theta_0\|^2}\pi_0(\theta)d\theta \\
\geq &\pi_{0,\inf}\int_{B_{\epsilon'}} e^{\sum_n f_n(\theta)-f_n(\theta_0) - \gamma N r(\epsilon')\|\theta-\theta_0\|^2}d\theta \\
\geq &\pi_{0,\inf}e^{-t_a(\epsilon')}\int_{B_{\epsilon'}} 1_{\sum_n f_n(\theta)-f_n(\theta_0) \geq N(\bar{f}(\theta)-\bar{f}(\theta_0)) - t_a} e^{-L_1'(\epsilon')N\|\theta-\theta_0\|^2}d\theta
\end{aligned} \tag{36}$$

$$\text{where} \quad \pi_{0,\inf} \coloneqq \inf_{\theta\in B_{\epsilon'}} \pi_0(\theta), \quad L_1'(\epsilon') \coloneqq L_1 + \gamma r(\epsilon'), \quad \text{and} \quad t_a(\epsilon') \coloneqq \left(\frac{DL_2}{L_1'(\epsilon')a}\right)^{1/2} \text{ for } a > 0,$$

where the last line follows from assumption **A3** and the definition of $L_1'(\epsilon')$. The next goal is to bound Eq. (36) from below by

$$\geq \frac{\pi_{0,\inf}e^{-t_a(\epsilon')}}{2}\int_{B_{\epsilon'}} e^{-L_1'(\epsilon')N\|\theta-\theta_0\|^2}d\theta \tag{37}$$

with high probability. To begin, by Markov's inequality,

$$\begin{aligned}
&\Pr\left(\int_{B_{\epsilon'}} 1_{\sum_n f_n(\theta)-f_n(\theta_0) \geq N(\bar{f}(\theta)-\bar{f}(\theta_0)) - t_a(\epsilon')} e^{-L_1'(\epsilon')N\|\theta-\theta_0\|^2}d\theta < \frac{\int_{B_{\epsilon'}} e^{-L_1'(\epsilon')N\|\theta-\theta_0\|^2}d\theta}{2}\right) \\
&= \Pr\left(\int_{B_{\epsilon'}} 1_{\sum_n f_n(\theta)-f_n(\theta_0) < N(\bar{f}(\theta)-\bar{f}(\theta_0)) - t_a(\epsilon')} e^{-L_1'(\epsilon')N\|\theta-\theta_0\|^2}d\theta > \frac{\int_{B_{\epsilon'}} e^{-L_1'(\epsilon')N\|\theta-\theta_0\|^2}d\theta}{2}\right) \\
&\leq \frac{2\int_{B_{\epsilon'}} \Pr\left(\sum_n f_n(\theta) - f_n(\theta_0) < N(\bar{f}(\theta)-\bar{f}(\theta_0)) - t_a(\epsilon')\right) e^{-L_1'(\epsilon')N\|\theta-\theta_0\|^2}d\theta}{\int_{B_{\epsilon'}} e^{-L_1'(\epsilon')N\|\theta-\theta_0\|^2}d\theta}.
\end{aligned} \tag{38}$$

The first equality here follows from the fact that

$$\begin{aligned}
&1_{\sum_n f_n(\theta)-f_n(\theta_0) \geq N(\bar{f}(\theta)-\bar{f}(\theta_0)) - t_a(\epsilon')} e^{-L_1'(\epsilon')N\|\theta-\theta_0\|^2} \\
&+ 1_{\sum_n f_n(\theta)-f_n(\theta_0) < N(\bar{f}(\theta)-\bar{f}(\theta_0)) - t_a(\epsilon')} e^{-L_1'(\epsilon')N\|\theta-\theta_0\|^2} = e^{-L_1'(\epsilon')N\|\theta-\theta_0\|^2}.
\end{aligned}$$

Thus we can see that if the integral of one of the terms on the left hand side of the above is less than half the integral of the right hand side, then the integral of the other must be greater than half the integral of the right hand side.

We can apply Markov's inequality again to bound the probability inside the integral,

$$\begin{aligned}
&\Pr\left(\sum_n f_n(\theta) - f_n(\theta_0) < N(\bar{f}(\theta) - \bar{f}(\theta_0)) - t_a(\epsilon')\right) \\
=&\Pr\left(\frac{1}{N}\sum_n f_n(\theta_0) - f_n(\theta) - \mathbb{E}_p\left(\frac{1}{N}\sum_n f_n(\theta_0) - f_n(\theta)\right) > \frac{t_a(\epsilon')}{N}\right) \\
\leq&\frac{N^2}{t_a(\epsilon')^2}\mathrm{Var}_p\left(\frac{1}{N}\sum_n f_n(\theta_0) - f_n(\theta)\right) \\
=&\frac{N}{t_a(\epsilon')^2}\mathrm{Var}_p\left(f_1(\theta_0) - f_1(\theta)\right) \\
\leq&\frac{N}{t_a(\epsilon')^2}L_2\|\theta - \theta_0\|^2,
\end{aligned}$$

using assumption **A3**. Substituting this into Eq. (38) yields

$$\begin{aligned}
&\frac{2\int_{B_{\epsilon'}}\Pr\left(\sum_n f_n(\theta) - f_n(\theta_0) < N(\bar{f}(\theta) - \bar{f}(\theta_0)) - t_a(\epsilon')\right)e^{-L_1'(\epsilon')N\|\theta-\theta_0\|^2}d\theta}{\int_{B_{\epsilon'}}e^{-L_1'(\epsilon')N\|\theta-\theta_0\|^2}d\theta} \\
\leq&\frac{2L_2\int_{B_{\epsilon'}}N\|\theta-\theta_0\|^2e^{-L_1'(\epsilon')N\|\theta-\theta_0\|^2}d\theta}{t_a(\epsilon')^2\int_{B_{\epsilon'}}e^{-L_1'(\epsilon')N\|\theta-\theta_0\|^2}d\theta} \\
=&\frac{L_2}{L_1'(\epsilon')t_a(\epsilon')^2}\frac{\int_{\|u\|\leq\epsilon'\sqrt{2NL_1'(\epsilon')}}\|u\|^2e^{-\|u\|^2/2}du}{\int_{\|u\|\leq\epsilon'\sqrt{2NL_1'(\epsilon')}}e^{-\|u\|^2/2}du}
\end{aligned} \tag{39}$$

where we make the substitution $u = \sqrt{2NL_1'(\epsilon')}(\theta-\theta_0)$, and cancel the Jacobian terms in numerator and denominator. If $\epsilon' \geq \sqrt{\frac{D}{2NL_1'(\epsilon')}}$, the integral in the denominator can be bounded below by

$$\begin{aligned}
\int_{\|u\|\leq\epsilon'\sqrt{2NL_1'(\epsilon')}}e^{-\|u\|^2/2}du &\geq (2\pi)^{D/2}\int_{\|u\|^2\leq D}(2\pi)^{-D/2}e^{-\|u\|^2/2}du \\
&= (2\pi)^{D/2}\Pr\left(Y \leq D\right), \qquad Y \sim \chi^2_D \\
&\geq \frac{1}{2}(2\pi)^{D/2},
\end{aligned}$$

since the median of a χ^2_k-distribution is $\leq k$. Since $\Theta \subseteq \mathbb{R}^D$, we can upper bound the integral in the numerator by

$$\int_{\|u\|\leq\epsilon'\sqrt{2NL_1'(\epsilon')}}\|u\|^2e^{-\|u\|^2/2}du \leq (2\pi)^{D/2}\int\|u\|^2(2\pi)^{-D/2}e^{-\|u\|^2/2}du = D(2\pi)^{D/2},$$

and thus Eq. (39) can be upper bounded by

$$\frac{L_2}{L_1'(\epsilon')t_a(\epsilon')^2}\frac{\int_{\|u\|\leq\epsilon'\sqrt{2NL_1'(\epsilon')}}\|u\|^2e^{-\|u\|^2/2}du}{\int_{\|u\|\leq\epsilon'\sqrt{2NL_1'(\epsilon')}}e^{-\|u\|^2/2}du} \leq \frac{2L_2D}{L_1'(\epsilon')t_a(\epsilon')^2} = 2a,$$

using the definition of $t_a(\epsilon')$. Thus with probability at least $1-2a$, we can bound Eq. (36) from below by

$$\begin{aligned}\text{Eq. (36)} &\geq \frac{\pi_{0,\inf} e^{-t_a(\epsilon')}}{2} \int_{B_{\epsilon'}} e^{-L_1'(\epsilon')N\|\theta-\theta_0\|^2} d\theta \\ &\geq \frac{\pi_{0,\inf} e^{-t_a(\epsilon')}}{2} \left(\frac{1}{2L_1'(\epsilon')N}\right)^{D/2} \int_{\|u\|\leq \epsilon'\sqrt{2L_1'(\epsilon')N}} e^{-\|u\|^2/2} d\theta \\ &\geq \frac{\pi_{0,\inf} e^{-t_a(\epsilon')}}{4} \left(\frac{\pi}{L_1'(\epsilon')N}\right)^{D/2}, \end{aligned} \tag{40}$$

using the same substitution and χ^2 bound as before. This gives us our final lower bound on the denominator of $I_{\epsilon,1}$. To summarize the result at this point: assuming that

$$\sum_m v_m \nabla_\theta f_m(\theta_0) = \frac{S_N(\theta_0)}{\sqrt{N}} \tag{41}$$

$$\sum_m v_m (f_m^{(1)}(\theta) - f_m^{(1)}(\theta_0)) = \frac{\sum_n (f_n^{(1)}(\theta) - f_n^{(1)}(\theta_0))}{N} \tag{42}$$

$$\left|\sum_m v_m (R(x_m) - \mathbb{E}_p(R(X))\right| \leq 1, \tag{43}$$

we have that for any $a > 0$, $\sqrt{\frac{D}{2NL_1'(\epsilon')}} \leq \epsilon' < \epsilon_0$, and $0 \leq \epsilon < \epsilon_0$,

$$\begin{aligned} I_{\epsilon,1} &\leq \frac{e^{-NL_0\epsilon^2/8} \int_\Theta \mathbb{E}_p(\Delta_1 - \bar{\Delta})^2(\theta)\pi_0(\theta)d\theta}{\int_\Theta e^{t\sum_n (f_n(\theta)-f_n(\theta_0)) + N(1-t)\sum_m v_m(f_m(\theta)-f_m(\theta_0))}\pi_0(\theta)d\theta} \\ &\leq N^{D/2} e^{t_a(\epsilon') - NL_0\epsilon^2/8} \left(\frac{4}{\pi_{0,\inf}} \int_\Theta \mathbb{E}_p(\Delta_1 - \bar{\Delta})^2(\theta)\pi_0(\theta)d\theta \left(\frac{L_1'(\epsilon')}{\pi}\right)^{D/2}\right), \end{aligned} \tag{44}$$

with probability at least

$$1 - 4Me^{-NL_0\epsilon^2/4} - 3\frac{\mathbb{E}_p\left(\|\nabla_\theta f_1(\theta_0)\|^2\right)}{\epsilon^2 NL_0^2} - \frac{\mathrm{Var}_p(R(X))}{N} - 2a.$$

We now bound $I_{\epsilon,2}$, corresponding to the integral over B_ϵ. Throughout we assume that the same set of events occur as used in the analysis of $I_{\epsilon,1}$. Note that the denominator of $I_{\epsilon,2}$ is the same as in $I_{\epsilon,1}$, so we apply the same bound; here we focus on the numerator. Since $r(\epsilon)\gamma \leq L_0/2$ by **A2**, we can obtain the following bound for all $\theta \in B_\epsilon$:

$$\begin{aligned} &t\sum_n (f_n(\theta) - f_n(\theta_0)) + N(1-t)\sum_m v_m(f_m(\theta) - f_m(\theta_0)) \\ &\qquad \leq \sum_n (f_n(\theta) - f_n(\theta_0)) + N(1-t)r(\epsilon)\|\theta-\theta_0\|^2\gamma \\ &\qquad \leq \sqrt{N}S_N(\theta_0)^T(\theta-\theta_0) - NL_0\|\theta-\theta_0\|^2 + N(1-t)r(\epsilon)\|\theta-\theta_0\|^2\gamma \\ &\qquad \leq \sqrt{N}S_N(\theta_0)^T(\theta-\theta_0) - \frac{NL_0}{2}\|\theta-\theta_0\|^2 \\ &\qquad = -\frac{NL_0}{2}\left(\|\theta-\theta_0\|^2 - \frac{2S_N(\theta_0)^T(\theta-\theta_0)}{\sqrt{N}L_0}\right) \\ &\qquad \leq \frac{\|S_N(\theta_0)\|^2}{2L_0} - \frac{NL_0}{2}\left\|\theta - \theta_0 - \frac{S_N(\theta_0)}{\sqrt{N}L_0}\right\|^2. \end{aligned}$$

As before we write, for $\theta \in B_\epsilon$,

$$\frac{\sum_n f_n(\theta) - f_n(\theta_0)}{N} - \sum_m v_m(f_m(\theta) - f_m(\theta_0)) = \frac{\sum_n f_n^{(2)}(\theta) - f_n^{(2)}(\theta_0)}{N} - \sum_m v_m(f_m^{(2)}(\theta) - f_m^{(2)}(\theta_0)),$$

so the right hand side is bounded above by

$$r(\epsilon)\|\theta-\theta_0\|^2\left[2\mathbb{E}_p(R(X))+\frac{|\sum_n(R(x_n)-\mathbb{E}_p(R(X))|}{N}+\left|\sum_m v_m((R(x_m)-\mathbb{E}_p(R(X)))\right|\right]$$
$$\leq r(\epsilon)\|\theta-\theta_0\|^2\gamma,$$

and therefore

$$I_{\epsilon,2}\leq\frac{4r(\epsilon)^2\gamma^2e^{t_a(\epsilon')}(L_1'(\epsilon')N)^{D/2}\int_{B_\epsilon}\|\theta-\theta_0\|^4e^{\frac{\|S_N(\theta_0)\|^2}{2L_0}-\frac{NL_0}{2}\left\|\theta-\theta_0-\frac{S_N(\theta_0)}{\sqrt{N}L_0}\right\|^2}\pi_0(\theta)d\theta}{\pi^{D/2}\pi_{0,\inf}}.$$

We further bound the numerator using a similar technique as in the analysis of $I_{\epsilon,1}$:

$$\begin{aligned}\|\theta-\theta_0\|^2&=\left(\|\theta-\theta_0\|-\frac{\|S_N(\theta_0)\|}{\sqrt{N}L_0}+\frac{\|S_N(\theta_0)\|}{\sqrt{N}L_0}\right)^2\\&\leq 2\left(\|\theta-\theta_0\|-\frac{\|S_N(\theta_0)\|}{\sqrt{N}L_0}\right)^2+2\left(\frac{\|S_N(\theta_0)\|}{\sqrt{N}L_0}\right)^2\\&\leq 2\left\|(\theta-\theta_0)-\frac{S_N(\theta_0)}{\sqrt{N}L_0}\right\|^2+2\left(\frac{\|S_N(\theta_0)\|}{\sqrt{N}L_0}\right)^2.\end{aligned}$$

Note that by Markov's inequality,

$$\forall\zeta>0,\quad\|S_N(\theta_0)\|^2\leq D/\zeta,$$

with probability $\geq 1-\zeta\mathbb{E}_p(\|\nabla_\theta f_1(\theta_0)\|^2)/D$. If we further restrict to this event, we have that

$$\begin{aligned}I_{\epsilon,2}&\leq\frac{4e^{t_a(\epsilon')}r(\epsilon)^2\gamma^2\pi_{0,\sup}(L_1'(\epsilon')N)^{D/2}}{\pi^{D/2}\pi_{0,\inf}}\int_{B_\epsilon}\|\theta-\theta_0\|^4e^{\frac{\|S_N(\theta_0)\|^2}{2L_0}-\frac{NL_0}{2}\left\|\theta-\theta_0-\frac{S_N(\theta_0)}{\sqrt{N}L_0}\right\|^2}d\theta\\&\leq\frac{16e^{t_a(\epsilon')}r(\epsilon)^2\gamma^2\pi_{0,\sup}(L_1'(\epsilon')N)^{D/2}e^{\frac{D}{2L_0\zeta}}}{\pi^{D/2}\pi_{0,\inf}}\\&\qquad\times\int_{B_\epsilon}\left(\left\|(\theta-\theta_0)-\frac{S_N(\theta_0)}{\sqrt{N}L_0}\right\|^2+\left(\frac{\|S_N(\theta_0)\|}{\sqrt{N}L_0}\right)^2\right)^2e^{-\frac{NL_0}{2}\left\|\theta-\theta_0-\frac{S_N(\theta_0)}{\sqrt{N}L_0}\right\|^2}d\theta\\&\leq\frac{16e^{t_a(\epsilon')}r(\epsilon)^2\gamma^2\pi_{0,\sup}(L_1'(\epsilon')N)^{D/2}e^{\frac{D}{2L_0\zeta}}(NL_0)^{-(D/2+2)}}{\pi^{D/2}\pi_{0,\inf}}\\&\qquad\times\int\left(\|u\|^4+\frac{2\|u\|^2\|S_N(\theta_0)\|^2}{L_0}+\frac{\|S_N(\theta_0)\|^4}{L_0^2}\right)e^{-\frac{1}{2}\|u\|^2}du\\&\leq\frac{16e^{t_a(\epsilon')}r(\epsilon)^2\gamma^2\pi_{0,\sup}(L_1'(\epsilon')N)^{D/2}e^{\frac{D}{2L_0\zeta}}(NL_0)^{-(D/2+2)}(2\pi)^{D/2}}{\pi^{D/2}\pi_{0,\inf}}\\&\qquad\times\int\left(\|u\|^4+\frac{2\|u\|^2D}{L_0\zeta}+\frac{D^2}{L_0^2\zeta^2}\right)(2\pi)^{-D/2}e^{-\frac{1}{2}\|u\|^2}du\\&\leq\frac{16e^{t_a(\epsilon')}D^2r(\epsilon)^2\gamma^2\pi_{0,\sup}(2L_1'(\epsilon')/L_0)^{D/2}e^{\frac{D}{2L_0\zeta}}}{N^2L_0^2\pi_{0,\inf}}\left[3+\frac{2}{L_0\zeta}+\frac{1}{L_0^2\zeta^2}\right]\\&\leq\frac{16e^{t_a(\epsilon')}D^2r(\epsilon)^2\gamma^2\pi_{0,\sup}(2L_1'(\epsilon')/L_0)^{D/2}e^{\frac{D}{2L_0\zeta}}}{\zeta^2N^2L_0^2\pi_{0,\inf}}\left[3+\frac{2}{L_0}+\frac{1}{L_0^2}\right],\end{aligned}\tag{45}$$

using the substitution $u=\sqrt{NL_0}\left((\theta-\theta_0)-S_N(\theta_0)/(\sqrt{N}L_0)\right)$. Noting that $t_a(\epsilon')=\Theta(\sqrt{1/a})$, we will set $\sqrt{a}=\zeta$. What remains is to lower bound the probability that the three conditions Eqs. (41) to (43) hold. Denoting by $\psi(x)\in\mathbb{R}^{d_1}$ the coordinates of $f^{(1)}(x,\cdot)-f^{(1)}(x,\theta_0)$ in an orthonormal basis of S_1, we define $Z(x)\in\mathbb{R}^d$, $d\leq d_1+D$, as

$$Z(x):=\begin{bmatrix}\psi(x)\\ \text{indep }(\nabla_\theta f(x;\theta_0))\end{bmatrix},\tag{46}$$

where indep $(\cdot)$ selects the linearly independent components of the $\nabla_\theta f(x;\theta_0)$ (considered as functions of x). Without loss of generality we can choose $R(x)$ to be linearly independent of $Z(x)$ as a function of x. In this case, the three conditions hold simultaneously if

$$\sum_m v_m \psi_j(x_m) = \frac{\sum_n \psi_j(x_n)}{N} \quad \forall j, \quad \sum_m v_m (R(x_m) - \mathbb{E}_p R(X)) = 0, \quad \sum_m v_m \nabla f_m(\theta_0) = \frac{\sum_n \nabla f_n(\theta_0)}{N}.$$

As in the proof of Theorem 4.1, we can derive that for $\delta \in (0,1]$, if $M \gtrsim J(\delta)^{-1} D(\log N + 1)$, where

$$J(\delta) = \inf_{\substack{a \in \mathbb{R}^{d+1}; \\ \|a\|=1}} \Pr\left(\left\langle a, [Z(X), R(X)]^T \right\rangle > \frac{2\mathbb{E}_p(\|Z(X)\|^2)}{\sqrt{N\delta}} \right), \tag{47}$$

then with probability greater than $1 - \delta - e^{-J(\delta)M/4}$, Eqs. (41) to (43) hold. We will set $\delta = \zeta^2$. Our particular choice of Z using independent components ensures that $J(\delta)$ is of full rank, and therefore is bounded from below by 0.

We can now obtain a final high probability bound on the KL divergence by combining the bounds on $I_{\epsilon,1}$ and $I_{\epsilon,2}$. In particular, we have that if $M \gtrsim J(\zeta^2)^{-1} D(\log N + 1)$, we have that for any $\zeta > 0$, $\sqrt{\frac{D}{2NL_1'(\epsilon')}} \leq \epsilon' < \epsilon_0$, and $0 \leq \epsilon < \epsilon_0$,

$$\begin{aligned} KL(\pi_{w(v)} || \pi) &\leq 4N^2 \int_0^1 (1-t)\left(I_{\epsilon,1} + I_{\epsilon,2}\right) \mathrm{d}t \\ &\leq 4N^{D/2+2} e^{\frac{1}{\zeta}\sqrt{\frac{DL_2}{L_1'(\epsilon')}} - NL_0\epsilon^2/8} \left(\frac{2}{\pi_{0,\inf}} \int_\Theta \mathbb{E}_p(\Delta_1 - \bar{\Delta})^2(\theta)\pi_0(\theta) d\theta \left(\frac{L_1'(\epsilon')}{\pi} \right)^{D/2} \right) \\ &\quad + \frac{4r(\epsilon)^2 e^{\frac{1}{\zeta}\left(\sqrt{\frac{DL_2}{L_1'(\epsilon')}} + \frac{D}{2L_0}\right)}}{\zeta^2} \left(\frac{8D^2\gamma^2\pi_{0,\sup}(2L_1'(\epsilon')/L_0)^{D/2}}{L_0^2\pi_{0,\inf}} \left[3 + \frac{2}{L_0} + \frac{1}{L_0^2} \right] \right) \end{aligned}$$

with probability at least

$$1 - e^{-J(\zeta^2)M/4} - 4Me^{-NL_0\epsilon^2/4} - 3\frac{\mathbb{E}_p\left(\|\nabla_\theta f_1(\theta_0)\|^2\right)}{\epsilon^2 NL_0^2} - \frac{\mathrm{Var}_p(R(X))}{N} - 3\zeta^2 - \zeta\frac{\mathbb{E}_p(\|\nabla_\theta f_1(\theta_0)\|^2)}{D}.$$

If $\zeta^{-1} = o(-\log r(\epsilon))$ as $\epsilon \to 0$, then the second term in the KL bound converges to 0. Consequently if $\log N + |\log r(\epsilon)| = o(N\epsilon^2)$ as $N \to \infty$, the KL divergence converges to 0 as $N \to \infty$. Similarly, $J(\zeta^2)$ converges to a nonzero constant as long as $\zeta\sqrt{N} \to \infty$. Both conditions are possible to satisfy, for any $r(\epsilon)$, by making $\epsilon \to 0$ slowly enough as a function of N.

□

B.3 Proof of Theorem 4.3

Proof. For step size $\gamma \in [0,1]$, the result $\hat{w}_{k+1}$ of the approximate Newton step prior to projection onto $\mathcal{W}$ is

$$\begin{aligned} \hat{w}_{k+1} &= w_k + \gamma(G(w_k) + \tau I)^{-1} H(w_k)(1 - w_k) \\ &= w_k + \gamma(G(w_k) + \tau I)^{-1} H(w_k)(w_k^\star - w_k) + \gamma(G(w_k) + \tau I)^{-1} H(w_k)(1 - w_k^\star) \\ &= w_k + \gamma(G(w_k) + \tau I)^{-1} G(w_k)(w_k^\star - w_k) + \gamma(G(w_k) + \tau I)^{-1} H(w_k)(1 - w_k^\star), \end{aligned}$$

where the last line follows from the fact that $w_k, w_k^\star \in \mathcal{W}$. Consider the distance $\left\|w_{k+1} - w_{k+1}^\star\right\|$. Since $w_{k+1}^\star$ is the projection of w_{k+1} onto a convex set $\mathcal{W}^\star$, we have that

$$\left\|w_{k+1} - w_{k+1}^\star\right\| \leq \left\|w_{k+1} - w_k^\star\right\|.$$

Furthermore, since w_{k+1} is the projection of $\hat{w}_{k+1}$ onto $\mathcal{W}$, and $w_k^\star \in \mathcal{W}$,

$$\left\|w_{k+1} - w_k^\star\right\| \leq \left\|\hat{w}_{k+1} - w_k^\star\right\|.$$

Then, we have that:

$$\begin{aligned}\|\hat{w}_{k+1} - w_k^\star\| &= \left\|\left(I - \gamma(G(w_k) + \tau I)^{-1} G(w_k)\right)(w_k - w_k^\star) + \gamma(G(w_k) + \tau I)^{-1} H(w_k)(1 - w_k^\star)\right\| \\ &\leq \left\|\left(I - \gamma(G(w_k) + \tau I)^{-1} G(w_k)\right)(w_k - w_k^\star)\right\| + \gamma(\epsilon\|w_k - w_k^\star\| + \delta) \\ &\leq \eta \|w_k - w_k^\star\| + \gamma\delta.\end{aligned}$$

The first bound follows from our assumption in the theorem statement, and the triangle inequality. The second bound follows via the following logic. Since $\pi_w \propto \exp(w^T f_{[M]}(\theta))\pi_0(\theta)$, it has the same support for all $w \in \mathcal{W}$. Since $G(w)$ is the covariance matrix of $f_{[M]}(\theta)$, the null space of $G(w)$ is spanned by the set of vectors $u \in \mathbb{R}^M$ such that $u^T f_{[M]}(\theta) = c$ holds π_w-almost everywhere, for some constant $c \in \mathbb{R}$. Therefore the null space of $G(w)$ is the same for all $w \in \mathcal{W}$. Since $w^T f_{[M]} = (w + u)^T f_{[M]}$ (up to a constant in θ) for any $w \in \mathcal{W}$ and u in the null space of $G(w)$, we can augment any closed convex subset $\mathcal{W}^\star \subseteq \arg\min_{w \in \mathcal{W}} \mathrm{KL}(\pi_w || \pi)$ with the null space of $G(w)$ to form $\mathcal{W}_2^\star = \{w + u : w \in \mathcal{W}^\star,\, u \in \operatorname{null} G(w), w + u \in \mathcal{W}\}$, which is still a closed convex subset of $\mathcal{W}^\star \subseteq \arg\min_{w \in \mathcal{W}} \mathrm{KL}(\pi_w || \pi)$. Therefore any maximal convex subset $\mathcal{W}^\star$ must be of the previous form augmented with the null space of $G(w)$. Because $w_k^\star$ is a projection, $w_k^\star - w_k$ must be orthogonal to this null space. Denoting the full eigendecomposition $G(w) = V\Lambda(w)V^T$, $V \in \mathbb{R}^{M \times M}$, $\Lambda(w) \in \mathbb{R}^{M \times M}$, including zero eigenvalues, and similarly the reduced eigendecomposition $G(w) = V_+\Lambda_+(w)V_+^T$ with the null space removed,

$$\begin{aligned}\left\|\left(I - \gamma(G(w_k) + \tau I)^{-1} G(w_k)\right)(w_k - w_k^\star)\right\| &= \left\|V\left(I - \gamma(\Lambda(w_k) + \tau)^{-1}\Lambda(w_k)\right)V^T(w_k - w_k^\star)\right\| \\ &= \left\|V_+\left(I - \gamma(\Lambda_+(w_k) + \tau)^{-1}\Lambda_+(w_k)\right)V_+^T(w_k - w_k^\star)\right\| \\ &\leq (1 - \gamma\xi)\left\|V_+^T(w_k - w_k^\star)\right\| \\ &= (1 - \gamma\xi)\left\|V^T(w_k - w_k^\star)\right\| \\ &= (1 - \gamma\xi)\left\|(w_k - w_k^\star)\right\|.\end{aligned}$$

Finally solving the earlier recursion,

$$\begin{aligned}\|w_k - w_k^\star\| &\leq \eta^k \|w_0 - w_0^\star\| + \gamma\delta \sum_{j=0}^{k-1} \eta^j \\ &= \eta^k \|w_0 - w_0^\star\| + \gamma\delta\left(\frac{1 - \eta^k}{1 - \eta}\right).\end{aligned}$$

□

C Experiments

In this section, we present some additional results, along with the full details of the Bayesian radial basis function regression experiment. We also discuss the use of a Laplace approximation for the low-cost approximation $\hat{\pi}$ needed for the sparse regression methods: greedy iterative geodesic ascent (GIGA) and iterative hard thresholding (IHT). Finally, we include a note on the overall performance of those methods.

C.1 Synthetic Gaussian location model

From Fig. 1 we see that our additional results match closely those in Section 5. Our method (QNC) consistently outperforms the other subsampling methods, while the Laplace approximation gives the best results, by design.

C.2 Bayesian sparse linear regression

From Fig. 2 we see that our additional results match closely those in Section 5. Our method (QNC) consistently outperforms the other methods. In order to assess the computation gains our coreset approach achieves, a useful metric is to calculate the number of posterior samples N_{sample} for which the time taken to obtain N_{sample} samples from the full posterior is the same as the time taken to construct a coreset using QNC, and then take N_{sample} samples from the coreset posterior. Here, we

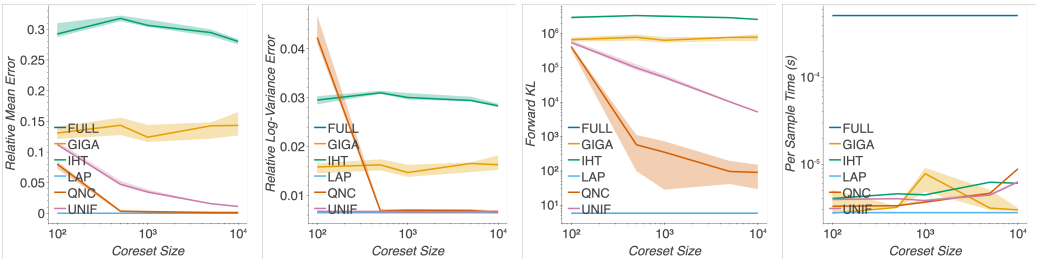


Figure 1: Relative mean and log-variance error, forward KL divergence and per sample time to sample from the respective posteriors for the synthetic Gaussian experiment. Our algorithm (QNC) consistently provides an improvement in coreset quality over the other subsampling methods. All methods provide a significant reduction in time to sample from the respective posteriors.

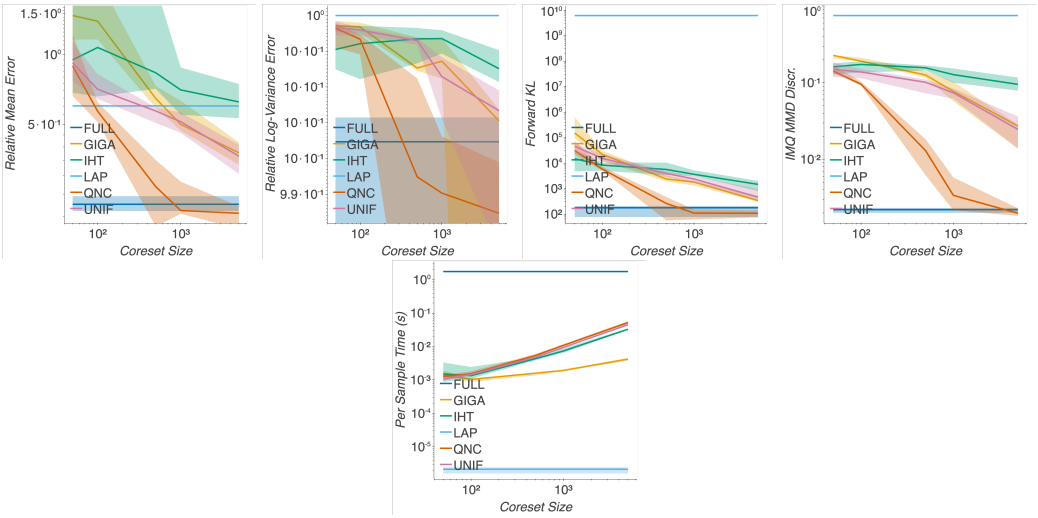


Figure 2: Relative mean and log-variance error, forward KL divergence, IMQ maximum mean discrepancy (MMD) and per sample time to sample from the respective posteriors for the sparse regression experiment. Our algorithm (QNC) consistently provides an improvement in coreset quality over the other methods. All methods provide a significant reduction in time to sample from the respective posteriors. The poor performance of the Laplace approximation in this heavy-tailed example can be seen particularly clearly in the forward KL and IMQ MMD plots.

find that $N_{sample} \approx 60$ for a coreset of size 1000. This is far fewer than you would ideally like to have in practice.

However, we note that these results are conservative, and in fact we expect that the gains from the coreset approach are more significant. This is because we perform sampling in each case using STAN[4], which uses C++. However, we construct the coresets in Python, and most of the coreset build time comes from the slowness of a non-compiled language. We expect that our estimated value of N_{sample} would be far smaller if we not only performed sampling in C++, but also constructed the coresets in C++.

In this experiment, the prior is heavy tailed, whilst the likelihood has sub-exponential tails. However, when the data is in fact concentrated on a particular subspace of the overall space, then the posterior can have heavy tails, even when the number of data points is large. We can clearly see this in Fig. 3, where we plot the quantiles of the (centred and scaled) marginal posterior distribution of the parameter λ_{13} to the quantiles of a standard normal distribution. This parameter is strictly positive, but we see that the right hand tail of the distribution is significantly heavier than a normal distribution.

In the forward KL and IMQ maximum mean discrepancy (MMD) plots, we can see clearly that the Laplace approximation is performing poorly, as it is underestimating the tails of the posterior distribution. Calculating the IMQ stein discrepancy in this experiment is computationally intractable with the size of dataset we consider.

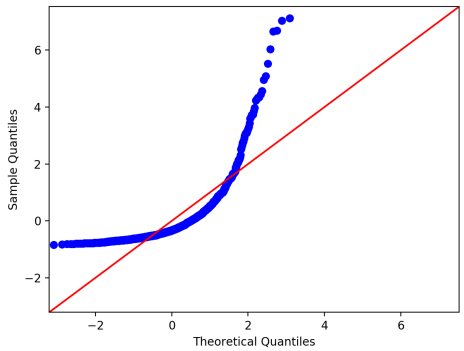


Figure 3: Quantile-quantile plot, comparing the quantiles of the (centred and scaled) marginal posterior distribution of λ_{13} to the quantiles of a standard normal distribution.

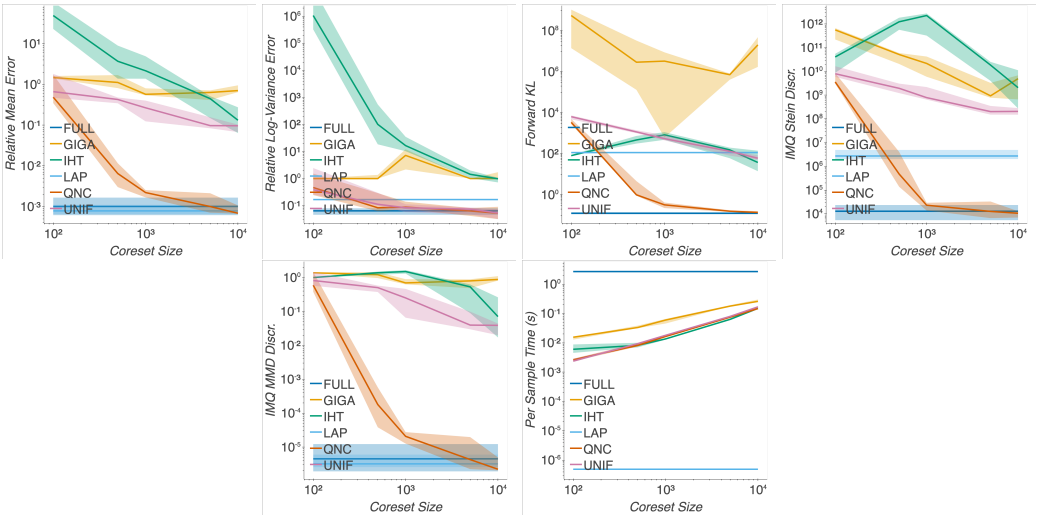


Figure 4: Relative mean and log-variance error, forward KL divergence, IMQ stein discrepancy, IMQ maximum mean discrepancy (MMD) and per sample time to sample from the respective posteriors for the logistic regression experiment. Our algorithm (QNC) generally provides an improvement in coreset quality over the other methods. All methods provide a significant reduction in time to sample from the respective posteriors.

C.3 Heavy-tailed Bayesian logistic regression

From Fig. 4 we see that our additional results match closely those in Section 5. Our method (QNC) consistently outperforms the other methods. As before, we can calculate the number of posterior samples N_{sample} for which the time taken to obtain N_{sample} samples from the full posterior is the same as the time taken to construct a coreset using QNC, and then take N_{sample} samples from the coreset posterior. Here, we find that $N_{sample} \approx 120$ for a coreset of size 1000. This is again far fewer than you would ideally like to have in practice.

C.4 Bayesian radial basis function regression

Our final comparison is on a Bayesian basis function regression example, and we provide the full details of that experiment here. We perform inference for the coefficients $\alpha \in \mathbb{R}^D$ in a linear combination of radial basis functions $b_k(x) = \exp\left(-1/2\sigma_k^2 \left(x - \mu_k\right)^2\right), k = 1, \ldots, D$,

$$y_n = b_n^T \alpha + \epsilon_n, \quad \epsilon_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(0, \sigma^2\right), \quad b_n = \left[\, b_1\left(x_n\right) \quad \cdots \quad b_D\left(x_n\right) \,\right]^T, \quad \alpha \sim \mathcal{N}\left(\mu_0, \sigma_0^2 I\right).$$

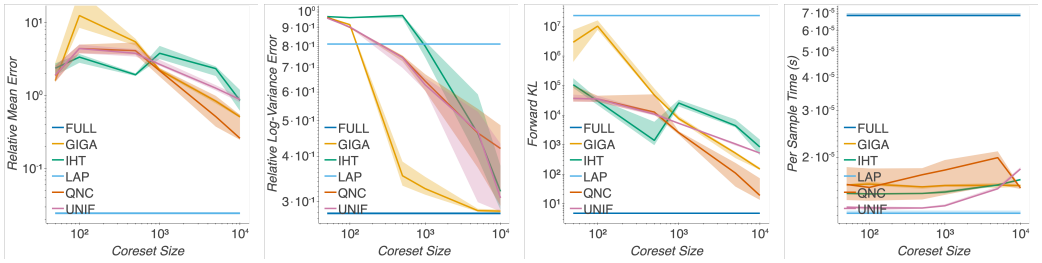


Figure 5: Relative mean and log-variance error, forward KL divergence and per sample time to sample from the respective posteriors for the basis regression experiment. Our algorithm (QNC) generally provides an improvement in coreset quality over the other methods. All methods provide a significant reduction in time to sample from the respective posteriors. The poor performance of the Laplace approximation in this high-dimensional example can be seen particularly clearly in the forward KL and log-variance plots.

The data consists of $N = 100,000$ records of house sale log-price $y_n \in \mathbb{R}$ as a function of latitude / longitude coordinates $x_n \in \mathbb{R}^2$ in the UK. [1] In order to perform inference we generate 50 basis functions for each of 6 scales $\sigma_k \in \{0.2, 0.4, 0.8, 1.2, 1.6, 2.0\}$ by generating means μ_k uniformly from the data, and including a basis with scale 100 (i.e. nearly constant) and mean corresponding to the mean latitude and longitude of the data. Thus, we have 301 total basis functions, and so $D = 301$. The prior and noise parameters $\mu_0, \sigma_0^2, \sigma^2$ are equal to the empirical mean, second moment, and variance of the price paid $(y_n)_{n=1}^N$ across the whole dataset, respectively. Closed form expressions are available for the subsampled posterior distributions [1, Appendix B], and we can sample from them without MCMC.

From Fig. 5 we see that our additional results match closely those in Section 5. Our method (QNC) generally outperforms the other methods, but has more trouble capturing the variance of the posterior, compared to the other subsampling approaches. Though the Laplace approximation captures the mean of the posterior well, it does not approximated the variance well, leading to its overall poor performance.

C.5 Comparison with Sparse Variational Inference

In this section, we provide a comparison of our method against SVI on a smaller dataset than that used in our original experiments. This experiment is the same as that in Section 5.1, except that we have reduced the dimension from 100 to 50, and the dataset size from $1,000,000$ to $10,000$. For SVI we use 100 optimization iterations, and only run 1 trial. From Fig. 6, we see that, even in this smaller data setting, the performance of SVI is not comparable. In reality, the number of optimization iterations needed is much higher than 100, which is reflected in the poor performance. However, we see that even for this number of optimization iterations the build time can be several orders of magnitude slower than any other method. Thus, we conclude that using SVI is not feasible on the size of datasets that we consider.

C.6 Sparse regression methods with Laplace approximation

In Section 5, we use a uniformly sampled coreset approximation of size M as the low-cost approximation $\hat{\pi}$ for GIGA and IHT. Here, we include additional results with a Laplace approximation used for $\hat{\pi}$. These are labeled as GIGA-LAP and IHT-LAP respectively. In Fig. 7, we see that in the synthetic Gaussian experiment, the choice of $\hat{\pi}$ has very little effect, and both GIGA and IHT perform quite badly in this large, high-dimensional example. In the sparse regression and basis function regression experiments, using a Laplace approximation for $\hat{\pi}$ leads to a worse performance. This may be unsurprising because the Laplace approximation (LAP) does not provide a good approximation to

[1] This dataset was constructed by merging housing prices from the UK land registry data `https://www.gov.uk/government/statistical-data-sets/price-paid-data-downloads` with latitude & longitude coordinates from the Geonames postal code data `http://download.geonames.org/export/zip/`. The housing price dataset contains HM Land Registry data © Crown copyright and database right 2021. This data is licensed under the Open Government Licence v3.0. The postal code data is licensed under a Creative Commons Attribution 4.0 License.

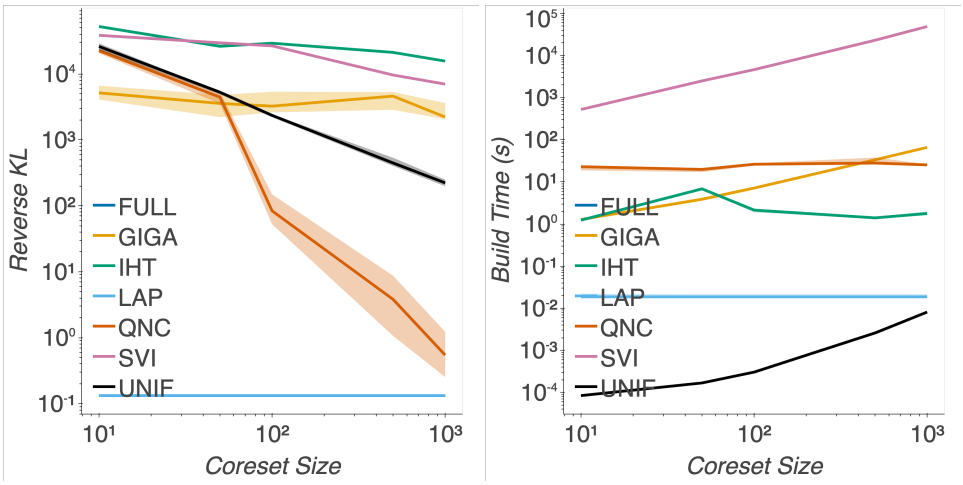
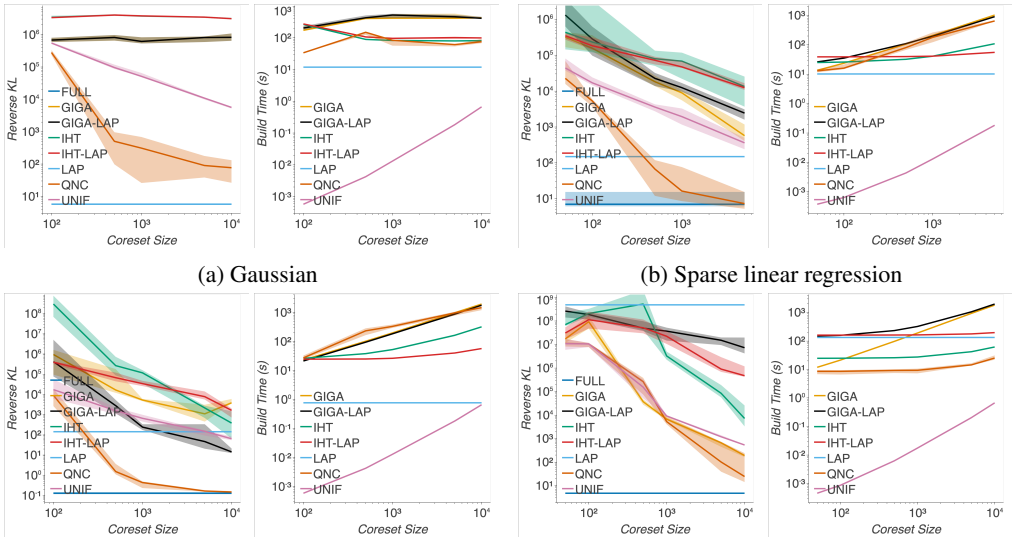


Figure 6: Reverse KL divergence (left) and build time in seconds (right) for the small synthetic Gaussian experiment.



(a) Gaussian

(b) Sparse linear regression

(c) Logistic regression

(d) Basis function regression

Figure 7: Reverse KL divergence (left) and build time in seconds (right) for each experiment, with added plots for GIGA and IHT with a Laplace approximation used for $\hat{\pi}$. We plot the median and a shaded area between the 25$^{\text{th}}$/75$^{\text{th}}$ percentiles over 10 random trials. Using a Laplace approximation used for $\hat{\pi}$ rarely provides an improvement in performance, and in some cases performs significantly worse than with the uniform choice of $\hat{\pi}$.

the true posterior in these settings, as detailed above and in Section 5. Only in the logistic regression experiment do we see an improvement in performance from GIGA-LAP over GIGA.

One reason for this is that the Laplace approximation provides limited help in coreset construction. Due to the large data set sizes used in our experiments, we only use a small number of samples ($S = 500$) in the coreset construction process. However, especially in high dimensions, we need a large number of samples to obtain a good L_2 approximation to the posterior from the Laplace approximation.

C.7 Performance of sparse regression methods

Throughout our experiments, we see that the sparse regression methods GIGA and IHT perform poorly. One reason for this is that these methods involve approximating a high dimensional integral with a small, fixed, number of samples. This problem persists no matter which choice of low-cost

approximation $\hat{\pi}$ we make, and is particularly problematic in high dimensions [5]. As noted by [1], methods like these are further limited by using a fixed approximation $\hat{\pi}$, rather than iteratively updating it as in the sparse variational inference approach.

C.8 Sensitivity Analysis

In this section we perform a sensitivity analysis for the parameters S, K_{tune} and τ that we use in Algorithm 1. We do this by repeating the Bayesian sparse linear regression experiment detailed in Section 5.2 for varying values of one of these parameters at a time (with all other parameters kept fixed). From Fig. 8 we see the following:

- **S:** In our original experiments, we use $S = 500$ throughout. From this analysis, we see that the performance of our method holds up for substantially lower values of S (though taking $S = 10$ is too low and leads to a degradation in performance). Moreover, doubling the number of samples does not lead to significantly better performance. The build time generally increases for larger S, as we would expect.
- τ : In our original experiments, we use $\tau = 0.01$ throughout. From this analysis, we see that taking too large a value of τ negatively affects the performance. This is in line with what we expect, since larger values skew the step away from a true Newton step. Our theory recommends taking τ as small as possible such that $G(w) + \tau I$ is still (numerically) invertible. Indeed, we see that decreasing τ can improve the performance of our model, though our choice still provides good results. Thus, we believe ours is a conservative choice that works well in practice.
- $\mathbf{K_{tune}}$: In our original experiments, we use $K_{tune} = 1$ throughout. In this analysis, we see that the choice of this parameter has very little effect on the performance of our method for this experiment. In our original experiments, we take $\gamma_k = 1$ for $k > K_{tune}$. When we perform our line search, our starting value is also $\gamma_k = 1$. We find that the line search stops immediately, meaning that for every value of K_{tune} we get essentially the same results.

  The purpose of this step in our algorithm is to guard against the case where our initial gradient and covariance estimates are very noisy, and we may want to take a smaller step initially. However, we see that this is in fact not needed for this experiment.

References

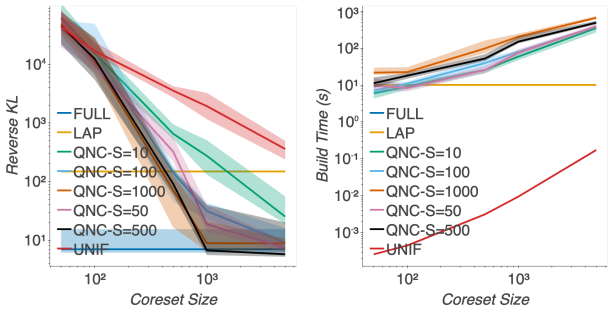
[1] Trevor Campbell and Boyan Beronov. Sparse variational inference: Bayesian coresets from scratch. In *Advances in Neural Information Processing Systems*, 2019.

[2] Shun-ichi Amari. *Information geometry and its applications*, volume 194. Springer, 2016.

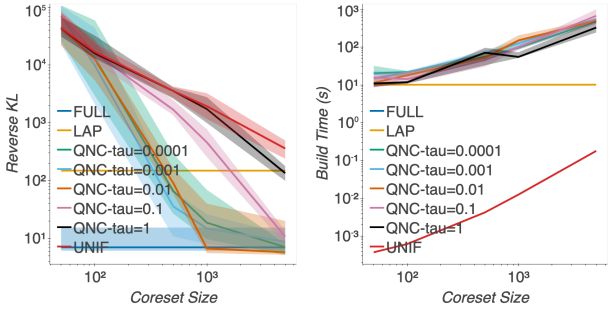
[3] Károly Böröczky and Gergely Wintsche. Covering the sphere by equal spherical balls. In Boris Aronov, Saugata Basu, János Pach, and Micha Sharir, editors, *Discrete and Computational Geometry*, volume 25 of *Algorithms and Combinatorics*, pages 235–251. Springer, 2003.

[4] Bob Carpenter, Andrew Gelman, Matthew D Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. Stan: A probabilistic programming language. *Journal of statistical software*, 76(1):1–32, 2017.

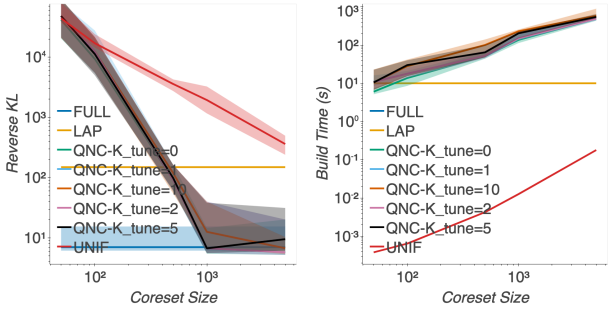
[5] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning*, volume 112. Springer, 2013.



(a) S



(b) τ



(c) K_{tune}

Figure 8: Sensitivity Analyses for the parameters (a) S, (b) τ and (c) K_{tune}. In each case, we plot the Reverse KL divergence (left) and build time in seconds (right) for the sparse regression experiment.