1 Introduction

Let $\{x_n\}_{n=1}^N \in \mathcal{X}$ be a dataset, $\theta \in \Theta \subset \mathbb{R}^D$ be a parameter, and $\pi_0(\theta)$ be a prior. Put $\mathcal{L}_n(\theta) = \log p(x_n|\theta)$ as a log-likelihood for nth observation and $\mathcal{L}(\theta) = \sum_{n=1}^N \mathcal{L}_n(\theta)$ as a log-likelihood. The true posterior $\pi(\theta)$ is given as

$$\pi(\theta) = \frac{1}{Z} \exp(\mathcal{L}(\theta))\pi_0(\theta),$$

where Z is the marginal likelihood: $Z = \int_\Theta \exp(\mathcal{L}(\theta))\pi(\theta)\, d\theta$.

For $\omega \in \mathbb{R}_+^N$, define $\mathcal{L}^\omega(\theta) = \sum_{n=1}^N \omega_n \mathcal{L}_n(\theta)$. The idea of Bayesian coreset is approximating $\mathcal{L}$ by using $\mathcal{L}^\omega$ with $\|\omega\|_0 \le M$ and $M \ll N$. Formally, the objective is

$$\text{minimize } \|\mathcal{L}^\omega - \mathcal{L}\|^2 \quad \text{sub. to } \|\omega\|_0 \le M.$$

2 Basic Algorithm from Huggins et al. (2016)

Algorithm 2.1 Coreset construction via importance sampling (Campbell and Broderick, 2017)

Require: $(\mathcal{L}_n)_{n=1}^N$, M, $\|\cdot\|$.
for $n \in \{1, 2, \dots, N\}$ **do**
  $\sigma_n \leftarrow \|\mathcal{L}_n\|$
end for
$\sigma \leftarrow \sum_{n=1}^N \sigma_n$
$(M_1, \dots, M_N) \sim \text{Multi}\left(M, \left(\frac{\sigma_n}{\sigma}\right)_{n=1}^N\right)$
for $n \in \{1, 2, \dots, N\}$ **do**
  $\omega_n \leftarrow \frac{\sigma}{\sigma_n}\frac{M_n}{M}$
end for
return ω

Definition 2.1 (Approximate dimension)**.** The *approximate dimension* $\dim(u_n)_{n=1}^N$ of N vectors in a normed vector space is the minimum value of $d \in \mathbb{N}$ such that all vectors u_n can be approximated using linear combinations of a set of d unit vectors $(v_j)_{j=1}^d$, $\|v_j\| = 1$:

$$\forall n \in \{1, \dots, N\}, \exists \alpha \in [-1, 1]^d : \left\| \frac{u_n}{\|u_n\|} - \sum_{j=1}^d \alpha_j v_j \right\| \le \frac{d}{\sqrt{N}}.$$

Theorem 2.1 (Campbell and Broderick, 2017)**.** *With probability* $\ge 1 - \delta$*, the output of the Algorithm* 2.1 *satisfies*

$$\|\mathcal{L}^\omega - \mathcal{L}\| \le \frac{\sigma}{\sqrt{M}} \left(2\dim(\mathcal{L}_n)_{n=1}^N + \bar{\eta}\sqrt{2\log\frac{1}{\delta}} \right), \quad \text{where } \bar{\eta} = \max_{n,m \in \{1,\dots,N\}} \left\| \frac{\mathcal{L}_n}{\sigma_n} - \frac{\mathcal{L}_m}{\sigma_m} \right\|.$$

Remark. The original theorem in the paper is *wrong*. See the remark in Lemma 2.1.

Lemma 2.1 (Campbell and Broderick, 2017)**.** *Suppose U and $\{U_m\}_{m=1}^M$ are i.i.d. random vectors in a normed vector space with discrete support on $\{u_n\}_{n=1}^N$ with probabilities $\{p_n\}_{n=1}^N$, and*

$$Y := \left\| \frac{1}{M} \sum_{m=1}^M U_m - \mathbb{E}[U] \right\|.$$

(*a*) *If* $\dim(u_n)_{n=1}^N \le d$ *where* $\dim$ *is given by Definition* 2.1*,*

$$\mathbb{E}[Y] \le \frac{d}{\sqrt{M}} \left(\sum_{n=1}^N \|u_n\| \sqrt{\frac{p_n(1-p_n)}{N}} + \sqrt{\mathbb{E}[\|U\|^2]} \right).$$

(*b*) *If the norm is a Hilbert norm,*

$$\mathbb{E}[Y] \le \frac{1}{\sqrt{M}} \sqrt{\mathbb{E}\left[\|U\|^2\right] - \|\mathbb{E}[U]\|^2}.$$

(*c*) *The random variable* $Y_m := \mathbb{E}[Y|\mathcal{F}_m]$ *with* $\mathcal{F}_m$ *the* σ*-algebra generated by* $U_1, \ldots, U_m$ *is a martingale that satisfies, for* $m \geq 1$*, both*

$$|Y_m - Y_{m-1}| \leq \frac{1}{M} \max_{n,l} \|u_n - u_l\|$$

and

$$\mathbb{E}\left[(Y_m - Y_{m-1})^2 | \mathcal{F}_{m-1}\right] \leq \frac{1}{M^2} \mathbb{E}\left[\|U - U_1\|^2\right]$$

almost surely.

Proof. (a) Denote $M_n = \sum_{m=1}^{M} \mathbb{I}(U_m = u_n)$. Also, denote α_n as the coefficients used to approximate u_n as in Definition 2.1. Then,

$$\begin{aligned}
\mathbb{E}[Y] &\leq \frac{1}{M}\mathbb{E}\left\|\sum_{n=1}^{N}(M_n - Mp_n)u_n\right\| \\
&\leq \frac{1}{M}\sum_{n=1}^{N}\mathbb{E}|M_n - Mp_n|\left\|u_n - \sum_{j=1}^{d}\alpha_{nj}\|u_n\|v_j\right\| + \frac{1}{M}\mathbb{E}\left\|\sum_{n=1}^{N}(M_n - Mp_n)\left(\sum_{j=1}^{d}\alpha_{nj}\|u_n\|v_j\right)\right\| \\
&\leq \frac{1}{M}\sum_{n=1}^{N}\frac{d\|u_n\|}{\sqrt{N}}\mathbb{E}|M_n - Mp_n| + \frac{1}{M}\sum_{j=1}^{d}\mathbb{E}\left|\sum_{n=1}^{N}(M_n - Mp_n)\|u_n\|\alpha_{nj}\right| \\
&\leq \frac{1}{M}\sum_{n=1}^{N}\frac{d\|u_n\|}{\sqrt{N}}\sqrt{\mathbb{E}(M_n - Mp_n)^2} + \frac{1}{M}\sum_{j=1}^{d}\sqrt{\mathbb{E}\left(\sum_{n=1}^{N}(M_n - Mp_n)\|u_n\|\alpha_{nj}\right)^2} \\
&\leq \frac{1}{\sqrt{M}}\sum_{n=1}^{N}d\|u_n\|\sqrt{\frac{p_n(1-p_n)}{N}} + \frac{1}{M}\sum_{j=1}^{d}\sqrt{\sum_{m=1}^{M}Var(A_{mj}\|U_{mj}\|)} \\
&= \frac{d}{\sqrt{M}}\left(\sum_{n=1}^{N}\|u_n\|\sqrt{\frac{p_n(1-p_n)}{N}} + \sqrt{\mathbb{E}[\|U\|^2]}\right),
\end{aligned}$$

where $A_{mj} = \sum_{n=1}^{N}\alpha_{nj}\mathbb{I}(U_m = u_n)$.

(b) Since $\|Z\|^2 = \langle Z, Z\rangle$,

$$\mathbb{E}[Y] \leq \sqrt{\mathbb{E}[Y^2]} = \frac{1}{M}\sqrt{\mathbb{E}\left\langle\sum_{m=1}^{M}(U_m - \mathbb{E}[U_m]), \sum_{m=1}^{M}(U_m - \mathbb{E}[U_m])\right\rangle} = \frac{1}{\sqrt{M}}\sqrt{\mathbb{E}\left[\|U\|^2\right] - \|\mathbb{E}[U]\|^2}.$$

(c) Trivially, $(Y_m)_{m=0}^{M}$ is a martingale. Fix $m \geq 1$, and put $U'_l = U_l$ for $l \neq m$ and $U'_m \stackrel{d}{=} U_m$ with $U'_m \perp U$ and $U'_m \perp U_l$ for all l. Then,

$$\begin{aligned}
|Y_m - Y_{m-1}| &= \left|\mathbb{E}\left[\left\|\frac{1}{M}\sum_{l=1}^{M}U_l - \mathbb{E}[U]\right\| \mid \mathcal{F}_m\right] - Y_{m-1}\right| \\
&\leq \left|\mathbb{E}\left[\left\|\frac{1}{M}(U_m - U'_m)\right\| \mid \mathcal{F}_m\right] + \mathbb{E}\left[\left\|\frac{1}{M}\sum_{l=1}^{M}U'_l - \mathbb{E}[U]\right\| \mid \mathcal{F}_m\right] - Y_{m-1}\right| \\
&= \frac{1}{M}\mathbb{E}\left[\|U_m - U'_m\| \mid \mathcal{F}_m\right] \leq \frac{1}{M}\max_{n,l}\|u_n - u_l\|,
\end{aligned}$$

$$\mathbb{E}\left[(Y_m - Y_{m-1})^2 \mid \mathcal{F}_{m-1}\right] \leq \mathbb{E}\left[\left(\frac{1}{M}\mathbb{E}\left[\|U_m - U'_m\| \mid \mathcal{F}_m\right]\right)^2 \mid \mathcal{F}_{m-1}\right] \leq \frac{1}{M^2}\mathbb{E}\left[\|U_m - U'_m\|^2\right]. \qquad \square$$

Remark. $Var\|U\|$ was in the original statement of Lemma 2.1(a) instead of $\mathbb{E}[\|U\|^2]$, which is trivially incorrect.

Proof of Theorem 2.1. Note that the conditions for Lemma 2.1 are satisfied by putting $u_n = \sigma\mathcal{L}_n/\sigma_n$, $p_n = \sigma_n/\sigma$, and $Y = \|\mathcal{L}^\omega - \mathcal{L}\|$. This implies that $|Y_m - Y_{m-1}| \leq \frac{\sigma\bar{\eta}}{M}$, so applying Azuma's inequality yields

$$Y \leq \mathbb{E}[Y] + \frac{\sigma\bar{\eta}}{\sqrt{M}}\sqrt{2\log\frac{1}{\delta}} \quad \text{with probability } \geq 1 - \delta.$$

By applying Lemma 2.1 again, we can obtain

$$
\begin{aligned}
Y &\leq \frac{\dim(\mathcal{L}_n)_{n=1}^N}{\sqrt{M}}\left(\|u_n\|\sum_{n=1}^{N}\sqrt{\frac{p_n(1-p_n)}{N}}+\sqrt{\mathbb{E}[\|U\|^2]}\right)+\frac{\sigma\bar{\eta}}{\sqrt{M}}\sqrt{2\log\frac{1}{\delta}} \\
&\leq \frac{\sigma}{\sqrt{M}}\left(2\dim(\mathcal{L}_n)_{n=1}^N+\bar{\eta}\sqrt{2\log\frac{1}{\delta}}\right) \quad \text{with probability } \geq 1-\delta.
\end{aligned}
$$

Note that $\|u_n\| = \sigma$ for all n. □

3 Using Hilbert Norm Gives More Efficient Result (Campbell and Broderick, 2017)

Campbell and Broderick (2017) suggested using a Hilbert norm (*i.e.*, a norm defined on inner product spaces) to incorporate with *directional* informations.

Theorem 3.1 (Campbell and Broderick, 2017)**.** *With probability* $\geq 1-\delta$*, the output of the Algorithm* 2.1 *satisfies*

$$
\|\mathcal{L}^{\omega}-\mathcal{L}\| \leq \frac{\sigma}{\sqrt{M}}\left(\eta+\eta_M\sqrt{2\log\frac{1}{\delta}}\right)
$$

where $\|\cdot\|$ *is a Hilbert norm and*

$$
\eta=\sqrt{1-\frac{\|\mathcal{L}\|^2}{\sigma^2}}, \quad \eta_M=\min\left\{\bar{\eta},\eta\sqrt{\frac{2M\eta^2}{\bar{\eta}^2\log\frac{1}{\delta}}}H^{-1}\left(\frac{\bar{\eta}^2\log\frac{1}{\delta}}{2M\eta^2}\right)\right\}, \quad H(y)=(1+y)\log(1+y)-y.
$$

Proof. Applying Azuma's inequality and martingale Bennet inequality gives

$$
Y \leq \mathbb{E}[Y]+\min\left\{\frac{\sigma\bar{\eta}}{\sqrt{M}}\sqrt{2\log\frac{1}{\delta}},\frac{2\sigma\eta^2}{\bar{\eta}}H^{-1}\left(\frac{\bar{\eta}^2}{2M\eta^2}\log\frac{1}{\delta}\right)\right\} \quad \text{with probability } \geq 1-\delta.
$$

By applying Lemma 2.1 again, we can obtain

$$
Y \leq \frac{\sigma\eta}{\sqrt{M}}+\frac{\sigma\eta_M}{\sqrt{M}}\sqrt{2\log\frac{1}{\delta}}=\frac{\sigma}{\sqrt{M}}\left(\eta+\eta_M\sqrt{2\log\frac{1}{\delta}}\right) \quad \text{with probability } \geq 1-\delta.
$$ □

In addition, Campbell and Broderick (2017) made some relaxation on the original optimization problem, resulting in the following objective:

$$
\text{minimize } \|\mathcal{L}^{\omega}-\mathcal{L}\|^2 \quad \text{sub. to } \sum_{n=1}^{N}\sigma_n\omega_n=\sigma.
$$

They solve this problem by using Frank–Wolfe algorithm, which gives a more efficient result.

Theorem 3.2 (Campbell and Broderick, 2017)**.** *The output of the Algorithm* 3.1 *satisfies*

$$
\|\mathcal{L}^{\omega}-\mathcal{L}\| \leq \frac{\sigma\eta\bar{\eta}\nu}{\sqrt{\bar{\eta}^2\nu^{-2(M-2)}+\eta^2(M-1)}} \leq \frac{\sigma\bar{\eta}}{\sqrt{M}},
$$

where $\nu=\sqrt{1-r^2/\sigma^2\bar{\eta}^2}$ *and* r *is the distance from* $\mathcal{L}$ *to the nearest boundary of the convex hull of* $\{\sigma\mathcal{L}_n/\sigma_n\}_{n=1}^N$.

Proof. See the paper. □

4 The Most Recent Algorithm is Campbell and Broderick (2018)

Campbell and Broderick (2018) found that Campbell and Broderick (2017) underestimates posterior uncertainty, so they added a scale term in the objective:

$$
\text{minimize } \|\alpha\mathcal{L}^{\omega}-\mathcal{L}\|^2 \quad \text{sub. to } \alpha \geq 0, \|\omega\|_0 \leq M.
$$

Since α can be solved analytically, this results in

$$
\text{maximize } \langle \ell^{\omega},\ell\rangle \quad \text{sub. to } \|\ell^{\omega}\|=1, \|\omega\|_0 \leq M.
$$

Applying the greedy algorithm gives Algorithm 4.1.

Algorithm 3.1 Coreset construction via Frank–Wolfe (Campbell and Broderick, 2017)

Require: $(\mathcal{L}_n)_{n=1}^N$, M, $\langle \cdot, \cdot \rangle$.

for $n \in \{1, 2, \ldots, N\}$ **do**
  $\sigma_n \leftarrow \|\mathcal{L}_n\|$
end for
$\sigma \leftarrow \sum_{n=1}^N \sigma_n$
$m \leftarrow \arg\max_{n \in \{1,2,\ldots,N\}} \left\langle \mathcal{L}, \frac{1}{\sigma_n}\mathcal{L}_n \right\rangle$
$\omega \leftarrow \frac{\sigma}{\sigma_m}\mathbf{1}_m$
repeat
  $m \leftarrow \arg\max_{n \in \{1,2,\ldots,N\}} \left\langle \mathcal{L} - \mathcal{L}^\omega, \frac{1}{\sigma_n}\mathcal{L}_n \right\rangle$
  $\gamma \leftarrow \frac{\left\langle \mathcal{L}_m - \mathcal{L}^\omega, \frac{\sigma}{\sigma_m}\mathcal{L}_m - \mathcal{L}^\omega \right\rangle}{\|\frac{\sigma}{\sigma_m}\mathcal{L}_m - \mathcal{L}^\omega\|}$
  $\omega \leftarrow (1-\gamma)\omega + \gamma\frac{\sigma}{\sigma_m}\mathbf{1}_m$
until $M - 1$ times
return ω

Algorithm 4.1 GIGA: Greedy Iterative Geodesic Ascent (Campbell and Broderick, 2018)

Require: $(\mathcal{L}_n)_{n=1}^N$, M, $\langle \cdot, \cdot \rangle$.

for $n \in \{1, 2, \ldots, N\}$ **do**
  $\ell_n \leftarrow \frac{\mathcal{L}_n}{\|\mathcal{L}_n\|}$
end for
$\ell \leftarrow \frac{\mathcal{L}}{\|\mathcal{L}\|}$
$\omega \leftarrow \mathbf{0}$
repeat
  for $n \in \{1, 2, \ldots, N\}$ **do**
    $d_n \leftarrow \frac{\ell_n - \langle \ell_n, \ell^\omega \rangle \ell^\omega}{\|\ell_n - \langle \ell_n, \ell^\omega \rangle \ell^\omega\|}$
  end for
  $d \leftarrow \frac{\ell - \langle \ell, \ell^\omega \rangle \ell^\omega}{\|\ell - \langle \ell, \ell^\omega \rangle \ell^\omega\|}$
  $k \leftarrow \arg\max_{n \in \{1,2,\ldots,N\}} \langle d, d_n \rangle$
  $\xi_1 \leftarrow \langle \ell, \ell_k \rangle$,$\xi_2 \leftarrow \langle \ell, \ell^\omega \rangle$,$\xi_3 \leftarrow \langle \ell_k, \ell^\omega \rangle$
  $\gamma \leftarrow \frac{\xi_0 - \xi_1\xi_2}{(\xi_0 - \xi_1\xi_2) + (\xi_1 - \xi_0\xi_2)}$
  $\omega \leftarrow \frac{(1-\gamma)\omega + \gamma\mathbf{1}_k}{\|(1-\gamma)\ell^\omega + \gamma\ell_k\|}$
until M times
for $n \in \{1, 2, \ldots, N\}$ **do**
  $\omega_n \leftarrow \frac{\|\mathcal{L}\|}{\|\mathcal{L}_n\|}\langle \ell^\omega, \ell \rangle\, \omega_n$
end for
return ω

Theorem 4.1 (Campbell and Broderick, 2018)**.** *The output of the Algorithm* 4.1 *satisfies* $\|\mathcal{L}^\omega - \mathcal{L}\| \le \eta\|\mathcal{L}\|\nu_M$*, where* ν_M *is decreasing and* ≤ 1 *for all* $M \in \mathbb{N}$*,* $\nu_M = O(\nu^M)$ *for some* $0 < \nu < 1$*, and*

$$\eta = \sqrt{1 - \left(\max_{n \in \{1,\ldots,N\}} \left\langle \frac{\mathcal{L}_n}{\|\mathcal{L}_n\|}, \frac{\mathcal{L}}{\|\mathcal{L}\|} \right\rangle\right)^2}$$

Proof. See the paper. □

5 Random Projection

Which norm is the most suitable for picking the coreset? Campbell and Broderick (2017) suggested followings:

$$\begin{cases} \langle \mathcal{L}_n, \mathcal{L}_m \rangle_{\hat{\pi},F} = \mathbb{E}_{\hat{\pi}}\left[\nabla\mathcal{L}_n(\theta)^\top \nabla\mathcal{L}_m(\theta)\right], \\ \langle \mathcal{L}_n, \mathcal{L}_m \rangle_{\hat{\pi},2} = \mathbb{E}_{\hat{\pi}}\left[\mathcal{L}_n(\theta)\mathcal{L}_m(\theta)\right], \end{cases}$$

where $\hat{\pi}$ would ideally be chosen equal to π to emphasize discrepancies that are in regions of high posterior mass. Unfortunately, evaluating such norms is often intractable. So they suggested using random projections of the $(\mathcal{L}_n)_{n=1}^N$ into a J dimensional vector space using samples from $\hat{\pi}$ (see Algorithm 5.1).

Algorithm 5.1 Random projection (Campbell and Broderick, 2017)

Require: $(\mathcal{L}_n)_{n=1}^N$, $\hat{\pi}$, M, J.
for $j \in \{1, 2, \ldots, J\}$ **do**
  $\mu_j \sim_{i.i.d.} \hat{\pi}$ and $d_j \sim_{i.i.d.}$ Unif$(\{1, 2, \ldots, D\})$.
end for
for $n \in \{1, 2, \ldots, N\}$ **do**
  $\hat{\mathcal{L}}_n \leftarrow \sqrt{D/J}[(\nabla \mathcal{L}_n(\mu_1))_{d_1}, \ldots, (\nabla \mathcal{L}_n(\mu_J))_{d_J}]^\top$ or $\hat{\mathcal{L}}_n \leftarrow \sqrt{1/J}[\mathcal{L}_n(\mu_1), \ldots, \mathcal{L}_n(\mu_J)]^\top$
end for
return CoresetAlgorithm $\left((v_n)_{n=1}^N, M, \|\cdot\|_2\right)$

Theorem 5.1 (Campbell and Broderick, 2017). *Let $\mu \sim \hat{\pi}$, $d \sim$ Unif$(\{1, \ldots, D\})$, and suppose $D\nabla\mathcal{L}_n(\mu)_d \nabla\mathcal{L}_m(\mu)_d$ (given $\|\cdot\|_{\hat{\pi},F}$) or $\mathcal{L}_n(\mu)\mathcal{L}_m(\mu)$ (given $\|\cdot\|_{\hat{\pi},2}$) is sub-Gaussian with constant ξ^2. With probability $\geq 1-\delta$, the output of the Algorithm* 5.1 *satisfies*

$$\|\mathcal{L}^\omega - \mathcal{L}\|_{\hat{\pi},2/F}^2 \leq \|\hat{\mathcal{L}}^\omega - \hat{\mathcal{L}}\|_2^2 + \|\omega - 1\|_1^2 \sqrt{\frac{2\xi^2}{J} \log \frac{2N^2}{\delta}}.$$

Proof. Consider only $\|\cdot\| = \|\cdot\|_{\hat{\pi},F}$. Denote K, V as the kernel matrix defined by $K_{ij} = \left\langle \mathcal{L}_i, \mathcal{L}_j \right\rangle$ and $V_{ij} = \left\langle \hat{\mathcal{L}}_i, \hat{\mathcal{L}}_j \right\rangle$. By Hoeffding's inequality,

$$\begin{aligned}
P\left(\max_{m,n} |K_{mn} - V_{mn}| \geq \epsilon\right) &\leq N^2 \max_{m,n} P(|K_{mn} - V_{mn}| \geq \epsilon) \\
&= N^2 \max_{m,n} P\left(\left|\sum_{j=1}^J \left(D\nabla\mathcal{L}_m(\mu_j)_{d_j} \nabla\mathcal{L}_n(\mu_j)_{d_j} - \mathbb{E}_{\hat{\pi}}\left[\nabla\mathcal{L}_m(\theta)^\top \nabla\mathcal{L}_n(\theta)\right]\right)\right| \geq J\epsilon\right) \\
&\leq 2N^2 \exp\left(-\frac{J\epsilon^2}{2\xi^2}\right).
\end{aligned}$$

This implies that

$$\max_{m,n} |K_{mn} - V_{mn}| \leq \sqrt{\frac{2\xi^2}{J} \log \frac{2N^2}{\delta}} \quad \text{with probability } \geq 1-\delta.$$

Therefore,

$$\begin{aligned}
\|\mathcal{L}^\omega - \mathcal{L}\|_{\hat{\pi},F}^2 - \|\hat{\mathcal{L}}^\omega - \hat{\mathcal{L}}\|_2^2 &= (\omega-1)^\top K(\omega-1) - (\omega-1)^\top V(\omega-1) \leq \sum_{m,n} |\omega_m - 1||\omega_n - 1||K_{mn} - V_{mn}| \\
&\leq \|\omega - 1\|_1^2 \max_{m,n} |K_{mn} - V_{mn}| \leq \|\omega - 1\|_1^2 \sqrt{\frac{2\xi^2}{J} \log \frac{2N^2}{\delta}} \quad \text{with probability } \geq 1-\delta.
\end{aligned}$$

The theorem can be proved for $\|\cdot\| = \|\cdot\|_{\hat{\pi},2}$ in a similiar manner. □

Bibliography

Campbell, T. and T. Broderick (2017). Automated Scalable Bayesian Inference via Hilbert Coresets. *arXiv preprint arXiv:1710.05053v1*.

Campbell, T. and T. Broderick (2018). Bayesian Coreset Construction via Greedy Iterative Geodesic Ascent. *arXiv preprint arXiv:1802.01737v2*.

Huggins, J. H., T. Campbell, and T. Broderick (2016). Coresets for Scalable Bayesian Logistic Regression. *arXiv preprint arXiv:1605.06423v3*.

A Supplementary Lemmas

Lemma A.1 (Azuma's inequality)**.** *Suppose* $(Y_m)_{m=0}^M$ *is a martingale adapted to the filtration* $(\mathcal{F}_m)_{m=0}^M$*. If there is a constant* ξ *such that for each* $m \in \{1, \dots, M\}$*,*

$$|Y_m - Y_{m-1}| \le \xi \quad a.s.,$$

then for all $\epsilon \ge 0$*,*

$$P(Y_M - Y_0 > \epsilon) \le e^{-\frac{\epsilon^2}{2M\xi^2}}.$$

Lemma A.2 (Martingale Bennet inequality)**.** *Suppose* $(Y_m)_{m=0}^M$ *is a martingale adapted to the filtration* $(\mathcal{F}_m)_{m=0}^M$*. If there are constants* ξ *and* τ^2 *such that for each* $m \in \{1, \dots, M\}$*,*

$$|Y_m - Y_{m-1}| \le \xi \quad and \quad \mathbb{E}\left[(Y_m - Y_{m-1})^2 | \mathcal{F}_{m-1}\right] \le \tau^2 \quad a.s.,$$

then for all $\epsilon \ge 0$*,*

$$P(Y_M - Y_0 > \epsilon) \le e^{-\frac{M\tau^2}{\xi^2} H\left(\frac{\epsilon\xi}{M\tau^2}\right)}, \quad where\ H(x) = (1+x)\log(1+x) - x.$$

Lemma A.3 (Hoeffding's inequality for sub-Gaussian)**.** *If* $(X_n)_{n=1}^N$ *are independent sub-Gaussian with constant* ξ_n^2 *respectively, then for all* $t \ge 0$*,*

$$P\left(\sum_{n=1}^N (X_n - \mathbb{E}X_n) \ge t\right) \le \exp\left(-\frac{t^2}{2\sum_{n=1}^N \xi_n^2}\right).$$