# Supplementary Material: Convergence Analysis of Block Coordinate Algorithms with Determinantal Sampling

## A PROOFS

### A.1 DPPs

*Proof of Theorem 1.* First, assume that $\alpha = 1$. Since $\mathbf{M} \succ \mathbf{0}$, we have $\det(\mathbf{M}_{SS}) > 0$ for all $S \subseteq [d]$. We will next use the following standard determinantal formula which holds for any $v \in \mathbb{R}^d$ and any invertible matrix $\mathbf{M}$:

$$\det(\mathbf{M})v^\top \mathbf{M}^{-1} v = \det(\mathbf{M} + vv^\top) - \det(\mathbf{M}). \tag{13}$$

Applying this formula to the submatrices of $\mathbf{M}$ and denoting by $v_S$ the sub-vector of $v$ indexed by $S$, we show that for any $v \in \mathbb{R}^d$:

$$\begin{aligned}
v^\top \mathbb{E}\big[(\mathbf{M}_S)^+\big] v &= \sum_{S\subseteq[d]} \frac{\det(\mathbf{M}_{SS})}{\det(\mathbf{I}+\mathbf{M})} v_S^\top \mathbf{M}_{SS}^{-1} v_S \\
&\overset{(13)}{=} \sum_{S\subseteq[d]} \frac{\det(\mathbf{M}_{SS} + v_S v_S^\top) - \det(\mathbf{M}_{SS})}{\det(\mathbf{I}+\mathbf{M})} \\
&= \frac{\sum_S \det([\mathbf{M}+vv^\top]_{SS}) - \sum_S \det(\mathbf{M}_{SS})}{\det(\mathbf{I}+\mathbf{M})} \\
&\overset{\text{(Lemma 1)}}{=} \frac{\det(\mathbf{I}+\mathbf{M}+vv^\top) - \det(\mathbf{I}+\mathbf{M})}{\det(\mathbf{I}+\mathbf{M})} \\
&\overset{(13)}{=} \frac{\det(\mathbf{I}+\mathbf{M})\, v^\top(\mathbf{I}+\mathbf{M})^{-1} v}{\det(\mathbf{I}+\mathbf{M})} \\
&= v^\top (\mathbf{I}+\mathbf{M})^{-1} v.
\end{aligned}$$

Since the above holds for all $v$, the equality also holds for the pd. matrices. To obtain the result with $\alpha \neq 1$, it suffices to replace $\mathbf{M}$ with $\frac{1}{\alpha}\mathbf{M}$. □

*Proof of Lemma 3.* The eigenvalues of $\mathbf{M}(\alpha\mathbf{I}+\mathbf{M})^{-1}$ are $\frac{\lambda_i}{\lambda_i+\alpha}$ so

$$\begin{aligned}
\mathbb{E}[|S|] &= \sum_{i=1}^d \frac{\lambda_i}{\lambda_i+\alpha} = \sum_{i=1}^d \frac{\lambda_i}{\lambda_i + \sum_{j\geq k}\lambda_j} \\
&= \sum_{i<k}^d \frac{\lambda_i}{\lambda_i + \sum_{j\geq k}\lambda_j} + \sum_{i\geq k}^d \frac{\lambda_i}{\lambda_i + \sum_{j\geq k}\lambda_j} \\
&< (k-1) + 1 = k,
\end{aligned}$$

which concludes the proof. □

### A.2 Convergence Analysis

*Proof of Theorem 2.*

$$\begin{aligned}
\sigma_1 &\overset{(7)}{=} \lambda_{\min}\left(\mathbf{M}^{1/2}\left(\alpha\mathbf{I}+\mathbf{M}\right)^{-1}\mathbf{M}^{1/2}\right) && (14)\\
&= \lambda_{min}\left(\left(\alpha\mathbf{M}^{-1}+\mathbf{I}\right)^{-1}\right) && (15)\\
&= \frac{1}{\lambda_{\max}\left(\alpha\mathbf{M}^{-1}+\mathbf{I}\right)} = \frac{1}{1+\alpha\lambda_{\max}(\mathbf{M}^{-1})} && (16)\\
&= \frac{\mu}{\mu+\alpha} && (17)\\
& && (18)
\end{aligned}$$

where $\mu = \lambda_{\min}(\mathbf{M})$. □

*Proof of Proposition 1.* By definition,

$$\frac{1}{\sigma(k+1)} = 1 + \frac{\sum_{i>k}^{d} \lambda_i}{\lambda_d} = 1 + \frac{\sum_{i>k-1}^{d} \lambda_i - \lambda_k}{\lambda_d} = \frac{1}{\sigma(k)} - \frac{\lambda_k}{\lambda_d}$$

Rearranging,

$$\frac{1}{\sigma(k)} = \frac{1}{\sigma(k+1)} + \frac{\lambda_k}{\lambda_d} \implies \sigma(k) = \frac{\sigma(k+1)\lambda_d}{\lambda_d + \lambda_k \sigma(k+1)}$$

Dividing the denominator and the numerator by $\lambda_d$ finishes the proof.

□

### A.3 Dual convergence rate

The dual convergence rate established in Qu et al. (2016) relies on the notion of expected separable over-approximation. Namely, the existence of $v \in \mathbb{R}^d$ s.t. $\mathbb{E}[\mathbf{M}_S] \preceq \mathbf{D}(p \circ v)$, where $p$ is the vector of marginal probabilities. In case of DPP sampling, one can choose $v = \text{diag}(\mathbf{M}) \circ \text{diag}(\mathbf{M}(\mathbf{M} + \alpha\mathbf{I})^{-1})^{-1}$, and apply dual convergence results established in this literature. By $\circ$ we denote element-wise product.

## B LEVERAGE SCORE SAMPLING VS DPP SAMPLING

We perform a simple experiment on the Gaussian Mixtures dataset where the matrix has a sparse spectrum. In Figure 5 we see that the optimization process is influenced minimally.

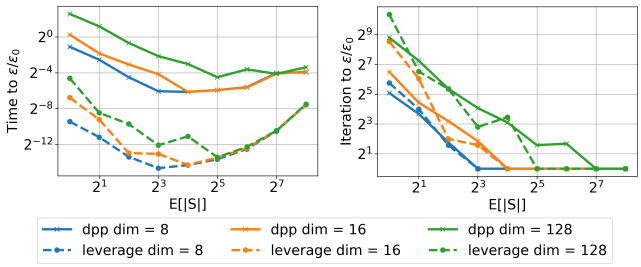


Figure 5: Comparison of leverage score sampling and DPP

## C RELATIVE SMOOTHNESS AND RELATIVE STRONG CONVEXITY

Recent works such as (Gower et al., 2019) and (Karimireddy et al., 2018) introduce the concepts of relative-smoothness, relative strong convexity and $c$-stability. These are weaker conditions than assumed in this paper. With these conditions, the proof techniques used to analyze coordinate descent algorithms are applicable to Newton-like algorithms, where instead of a fixed matrix $\mathbf{M}$, the actual Hessian $\mathbf{H}(x)$ can be used. The extension to $c-$stability is done trivially in Theorem 2 of Karimireddy et al. (2018), here we focus on a slightly more elaborate connection with relative smoothness and relative strong-convexity.

**Assumption 3** (Gower et al. (2019))**.** *There exists a constant $\tilde{L} \geq \tilde{\mu}$ such that for all $x, y \in \mathcal{Q} \subseteq \mathbb{R}^d$, where $\mathcal{Q} := \{x \in \mathbb{R}^d : f(x) \leq f(x_0)\}$:*

$$f(x) \leq f(y) + \langle \nabla f(y), x \rangle + \frac{\tilde{L}}{2} \|x - y\|_{\mathbf{H}(y)} \tag{19}$$

*and*

$$f(x) \geq f(y) + \langle \nabla f(y), x \rangle + \frac{\tilde{\mu}}{2} \|x - y\|_{\mathbf{H}(y)}. \tag{20}$$

Now the task is to analyze the algorithm with the following update rule, which is identical to general Newton rule when $S = [d]$,

$$x_{k+1} = x_k - \gamma(\mathbf{H}(x_k)_{S_k})^{+}\nabla f(x_k). \tag{21}$$

We fix a particular choice of $\gamma = 1/\tilde{L}$. This should be contrasted with the update rule (3).

Now given these assumption, we are able to show that the constant akin to $\sigma(\hat{S})$ appears in the analysis of this algorithm by utilizing the notions from (Gower et al., 2019). We sacrifice generality for the sake of brevity, and assume that range of $\mathbf{H}(x)$ spans whole $\mathbb{R}^d$ for each $x \in \mathcal{Q}$. Then, the following quantities resembling $\sigma(\hat{S})$ appear in the convergence analysis of the update rule (21)

$$\hat{\sigma}(\hat{S}, x) = \lambda_{\min}\left(\mathbb{E}_{\hat{S}}\left[\mathbf{H}^{1/2}(x)(\mathbf{H}(x)_{\hat{S}})^{+}\mathbf{H}(x)^{1/2}\right]\right) \tag{22}$$

and

$$\hat{\sigma}(\hat{S}) = \min_{x\in\mathcal{Q}} \hat{\sigma}(\hat{S}, x)$$

**Theorem 3** (Theorem 3.1 of Gower et al. (2019), modified)**.** *Let $f$ satisfy Assumption 3, and let $\mathbf{H}(x)$ be the Hessian at $x$ having range that spans whole $\mathbb{R}^d$ for all $x$. Then*

$$\mathbb{E}_{\hat{S}}[f(x_{k+1}) - f(x^*)] \leq \left(1 - \frac{\hat{\sigma}(\hat{S}, x_k)\mu}{L}\right)(f(x_k) - f(x^*),$$

*and*

$$\mathbb{E}_{\hat{S}}[f(x_k) - f(x^*)] \leq \left(1 - \frac{\hat{\sigma}(\hat{S})\mu}{L}\right)^{k}(f(x_0) - f(x^*),$$

*where $\hat{\sigma}(\hat{S}) = \min_{x\in\mathcal{Q}} \hat{\sigma}(\hat{S}, x)$ as in Equation (22).*

*Proof.* Minimizing the upper bound in (19) restricted to coordinates in $S_k$, we arrive at,

$$\begin{aligned}
f(x_{k+1}) - f(x_k) &\overset{(2)}{\leq} -\frac{1}{2\tilde{L}}\langle\nabla f(x_k), (\mathbf{H}(x_k)_{S_k})^{+}\nabla f(x_k)\rangle \\
\mathbb{E}[f(x_{k+1}) - f(x_k)] &\leq -\frac{1}{2\tilde{L}}\langle\nabla f(x_k), \mathbb{E}_{\hat{S}}[(\mathbf{H}(x_k)_{S_k})^{+}]\nabla f(x_k)\rangle \\
&\overset{(20),(22)}{\leq} -\frac{\mu}{L}\hat{\sigma}(\hat{S}, x)(f(x_k) - f(x^*)) \\
&\leq -\frac{\mu}{L}\hat{\sigma}(\hat{S})(f(x_k) - f(x^*))
\end{aligned}$$

rearranging finishes the proof. □

The following corollary states that with DPP sampling, the update rule in (21) can have a more interpretable convergence rate than stated in the Theorem 3.

**Corollary 1** (of Theorem 3)**.** *Under the assumption of Theorem 3, let additionally $S_k$ be a sample from sampling $\hat{S}_k \sim \mathrm{DPP}(\frac{1}{\alpha}\mathbf{H}(x_k))$, then*

$$\mathbb{E}_{\hat{S}_k}[f(x_{k+1}) - f(x^*)] \leq \left(1 - \left(\frac{\lambda(x_k)}{\lambda(x_k) + \alpha}\right)\frac{\mu}{L}\right)(f(x_k) - f(x^*),$$

*where $\lambda(x_k) = \lambda_{\min}(\mathbf{H}(x_k))$.*

The following lemma relates the complexity quantity defined above to the definition of $\sigma(\hat{S})$ used in the main body of this paper. Note that $\hat{\sigma}$ is larger than $\sigma$, even if the fixed over-approximation exists, as previously we assumed the over-approximation to be valid globally not just in $\mathcal{Q}$.

**Lemma 7.** *If for all $x \in \mathcal{Q}$, $\mathbf{M} \succeq \mathbf{H}(x) \succeq \kappa \mathbf{M} \succ 0$, then*

$$\hat{\sigma}(\hat{S}) \geq \kappa \sigma(\hat{S}).$$

*The relative smoothness, and strong-convexity can be chosen to be $\tilde{L} = 1$, and $\tilde{\mu} = 1$, respectively.*

*Proof.*

$$\begin{aligned}\hat{\sigma}(\hat{S}) &= \min_{x \in \mathcal{Q}} \min_{v \in \mathbb{R}^d} \frac{\langle v, \mathbb{E}_{\hat{S}} \left[\mathbf{H}^{1/2}(x)(\mathbf{H}(x)_{\hat{S}})^{+}\mathbf{H}(x)^{1/2}\right] v\rangle}{\|v\|_2^2} = \min_{v \in \mathbb{R}^d} \min_{x \in \mathcal{Q}} \frac{\langle v, \mathbb{E}_{\hat{S}} \left[\mathbf{H}^{1/2}(x)(\mathbf{H}(x)_{\hat{S}})^{+}\mathbf{H}(x)^{1/2}\right] v\rangle}{\|v\|_2^2} \\ &\geq \min_{v \in \mathbb{R}^d} \frac{\langle v, \mathbb{E}_{\hat{S}} \kappa \left[\mathbf{M}^{1/2}(\mathbf{M}_{\hat{S}})^{+}\mathbf{M}^{1/2}\right] v\rangle}{\|v\|_2^2} = \kappa \sigma(\hat{S})\end{aligned}$$

□

# D OTHER SAMPLINGS

The convergence properties of RNM with determinantal sampling depend solely on the spectral properties of $\mathbf{M}$. This is not true of other common samplings such as $\tau$-nice. Indeed we can improve or worsen the performance of $\tau$-nice sampling when $\mathbf{M}$ is transformed via spectrum preserving operation such as unitary transformation

$$\mathbf{M} \leftarrow \mathbf{R}^\top \mathbf{M} \mathbf{R}, \text{ where } \mathbf{R}^\top \mathbf{R} = \mathbf{I}.$$

Suppose that we are given an eigenvalues of the matrix $\mathbf{M}$, for any sampling $\hat{S}$ is it possible to find a spectrum preserving rotation such that $\sigma(\hat{S})$ is at least as small as $\sigma(\hat{S}_{\text{DPP}})$ which corresponds to DPP sampling with the same expected cardinality? The answer turns out to be negative, and we show counter-example.

**Remark 2** (Counter-example)**.** *Let $\hat{S}_1$ be a sampling such that $[n]$ is sampled with $1/2$ probability and $\emptyset$ and $1/2$ probability. The expected size of the subset $\mathbb{E}[|\hat{S}_1|] = d/2$ and $\sigma(\hat{S}_1) = \frac{1}{2}$ irrespective of the matrix $\mathbf{M}$.*

*Suppose matrix $\mathbf{M}$ has degenerate spectrum such that $\lambda$ is eigenvalue with multiplicity $d/2$ and $\mu$ is eigenvalue with $d/2$ multiplicity where $\lambda < \mu$. In order s.t. $\mathbb{E}[|S_{\text{DPP}}|] = \frac{d}{2}$, $\alpha = \sqrt{\lambda\mu}$, then $\sigma(\hat{S}_{\text{DPP}}) < \frac{1}{2}$.*

In what circumstances does DPP sampling perform better than a uniform sampling? First, we consider circumstances where uniform sampling is optimal.

## D.1 Uniform sampling

It is important to allow for variation in the off-diagonal of $\mathbf{M}$. If we consider only diagonal $\mathbf{M}$, the optimal sampling is uniform sampling.

**Lemma 8.** *Let $\mathbf{M}$ be diagonal. The quantity $\sigma(\hat{S})$ of a sampling over a power set $P([d])$ constrained by $\mathbb{E}[|\hat{S}|] = k$ is maximized for uniform samplings.*

*Proof of Lemma 8 .* We want to maximize the minimum eigenvalue of a matrix $\mathbf{M}^{1/2}\mathbb{E}[(\mathbf{M}_S)^{-1}]\mathbf{M}^{1/2}$. For a diagonal $\mathbf{M}$ we know that $(\mathbf{M}_S)^{-1} = (\mathbf{M}^{-1})_S$. Hence, $\mathbf{M}^{1/2}\mathbb{E}[(\mathbf{M}_S)^{-1}]\mathbf{M}^{1/2}\mathbf{D}(p)$, where $p$ is a vector of marginals $p_i = P(i \in \hat{S})$. Hence, the minimum eigenvalue is the minimum marginal probability subject to a constraint that $\mathbb{E}[|S|] = \sum_{j=1}^{d} P(j \in \hat{S}) \leq k$. This leads to an optimum where $P(i \in \hat{S}) = P(j \in \hat{S})$ for all $i, j \in [d]$. Hence the optimal sampling distribution is uniform. □

## D.2 Parallel Sampling

The parallel extension of the update method 3 has been considered in Mutný and Richtárik (2018) and Karimireddy et al. (2018). Namely, the authors consider a case, when the updates with $c$ machines are aggregated together to form a single update in the form $\approx \frac{1}{b}\sum_{j=1}^{c}(\mathbf{M}_{S_j})^{+}$, where $b$ is the aggregating parameter. It is known that for parallel disjoint samplings the convergence rate increases linearly with the number of processors. For independent samplings the aggregating parameter $b$ depends on the quantity,

$$\theta(\hat{S}) = \lambda_{\max}(\mathbf{M}^{1/2}\mathbb{E}[(\mathbf{M}_{\hat{S}})^{+}]\mathbf{M}^{1/2})$$

which in the case of DPP sampling is equal to $\theta = \frac{\lambda_1}{\lambda_1+\alpha}$. The quantity $\theta(\hat{S}) \in [\sigma(\hat{S}), 1]$, and as $\theta \to 1$, the aggregation operation becomes averaging $b \to c$. For DPP sampling, we can see an inverse relationship between increasing $\sigma(\hat{S})$ by increasing block size, which inherently makes the parallelization problem more difficult by increasing $\theta(\hat{S})$.