Convergence Analysis of Block Coordinate Algorithms with Determinantal Sampling

Mojmír Mutný[†]
Department of Computer Science
ETH Zurich, Switzerland
mmutny@inf.ethz.ch

Michał Dereziński[†]
Department of Statistics
University of California, Berkeley
mderezin@berkeley.edu

Andreas Krause
Department of Computer Science
ETH Zurich, Switzerland
krausea@inf.ethz.ch

Abstract

We analyze the convergence rate of the randomized Newton-like method introduced by Qu et al. (2016) for smooth and convex objectives, which uses random coordinate blocks of a Hessian-over-approximation matrix $\mathbf{M}$ instead of the true Hessian. The convergence analysis of the algorithm is challenging because of its complex dependence on the structure of $\mathbf{M}$. However, we show that when the coordinate blocks are sampled with probability proportional to their determinant, the convergence rate depends solely on the eigenvalue distribution of matrix $\mathbf{M}$, and has an analytically tractable form. To do so, we derive a fundamental new expectation formula for determinantal point processes. We show that determinantal sampling allows us to reason about the optimal subset size of blocks in terms of the spectrum of $\mathbf{M}$. Additionally, we provide a numerical evaluation of our analysis, demonstrating cases where determinantal sampling is superior or on par with uniform sampling.

1 INTRODUCTION

We study unconstrained optimization of the form:

$$\min_{x\in\mathbb{R}^d} f(x),$$

where we assume that the function $f : \mathbb{R}^d \to \mathbb{R}$ is smooth, convex, and potentially high dimensional. This problem commonly arises in empirical risk minimization (ERM, see Shalev-Shwartz and Ben-David, 2014). State-of-the-art approaches for minimization of convex ERM objectives with large numbers of data points include variants of stochastic gradient descent (SGD) such as SVRG (Johnson and Zhang, 2013), SARAH (Nguyen et al., 2017) and a plethora of others. Alternatively, one can approach the ERM problem via a dual formulation, where fast coordinate minimization techniques such as SDCA (Shalev-Shwartz and Zhang, 2013), or parallel coordinate descent (Richtárik and Takáč, 2016, 2015) can be applied. This is especially desirable in distributed and parallel environments (Richtárik and Takáč, 2016; Ma et al., 2015; Dünner et al., 2016). These approaches are closely related to methods that subsample the Hessian (Pilanci and Wainwright, 2015; Roosta-Khorasani and Mahoney, 2016; Roosta-Khorasani and Mahoney, 2016).

We study a block coordinate descent algorithm first introduced by Qu et al. (2016). In each iteration of this algorithm, we sample a block of coordinates and then solve a Newton step on the chosen coordinate subspace. However, in place of the true Hessian, a fixed over-approximation matrix $\mathbf{M}$ is used for the sake of efficiency. The Newton step is computed on a sparsified version of this matrix with all but the selected coordinates set to zero, denoted $\mathbf{M}_{\hat{S}}$ (see Section 1.2 for the complete notation). Originally, Qu et al. (2016) called this method Stochastic Dual Newton Ascent (SDNA), appealing to the fact that it operates in a dual ERM formulation. Later, it was also called a Stochastic Newton method (Mutný and Richtárik, 2018), while we use the name *Randomized Newton Method* (RNM) following Gower et al. (2019)*.

The sampling strategy for the coordinate blocks has a dramatic impact on the convergence rate (Qu and Richtárik, 2016). Gower and Richtárik (2015) demonstrate that by optimizing the sampling probabilities one can obtain very significant speedups, however this opti-

† Equal contribution. Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy. PMLR: Volume 108.

*Gower et al. (2019) consider a more general algorithm, relying on the novel assumptions of *relative* smoothness and convexity. We discuss this setting in Appendix C.

mization is a semidefinite program which may be even more challenging than the original optimization problem itself. Even when using a basic sampling strategy (such as uniform), the convergence analysis of RNM is challenging because it hinges on deriving the *expected pseudoinverse* of $\mathbf{M}_{\hat{S}}$, henceforth denoted $\mathbb{E}[(\mathbf{M}_{\hat{S}})^+]$. Prior to this work, no simple closed form expression was known for this quantity.

To overcome this challenge, we focus on a strategy of randomly sampling blocks of coordinates proportionally to the determinant of the corresponding submatrix of $\mathbf{M}$, which we call *determinantal sampling*. Similar sampling schemes have been analyzed in the context of stochastic optimization before (Zhang et al., 2017; Borsos et al., 2019). Recently, Rodomanov and Kropotov (2019) analyzed determinantal sampling for randomized block coordinate descent, however they imposed cardinality constraints on the block size, and as a result were unable to obtain a simple expression for $\mathbb{E}[(\mathbf{M}_{\hat{S}})^+]$.

We use determinantal sampling with randomized block size, which allows us to obtain a simple closed form expression for the expected pseudoinverse:

$$\mathbb{E}[(\mathbf{M}_{\hat{S}})^+] = (\alpha\mathbf{I} + \mathbf{M})^{-1},$$

where α is a tunable parameter that is used to control the expected block size. With the use of this new expectation formula, we establish novel bounds on the convergence rate of RNM depending on the spectral properties of the over-approximating matrix $\mathbf{M}$. For many instances of the problem, the matrix coincides with the data covariance, and spectral decays of such covariances are well understood (Blanchard et al., 2007). This allows us to predict the decay-specific behavior of RNM with determinantal sampling and recommend the optimal block size.

The cost of each iteration of RNM scales cubically with the size of the block due to matrix inversion. Qu et al. (2016) demonstrate numerically that for small blocks the optimization time decreases but at some point it starts to increase again. They surmise that the improvement is obtained only as long as the inversion cost is dominated by the other fixed per-iteration costs such as fetching from memory. However, whether the only possible speedup stems from this has remained unclear. We answer this question for determinantal sampling by deriving the optimal subset size in the case of kernel ridge regression. We show that when the eigenvalue decay is sufficiently rapid, then the gain in convergence rate can dominate the cost of inversion even for larger block sizes.

1.1 Contributions

The main contributions of this paper can be summarized as follows:

- We obtain a novel and remarkably simple expectation formula for determinantal sampling that allows us to derive a simple and closed form expression for the convergence rate of the Randomized Newton Method.
- This allows us to improve the previous bounds on the theoretical speedup of using coordinate blocks of larger sizes. For example, we show that in the case of kernel regression with a covariance operator that has exponentially decreasing spectrum, the theoretical speedup is *exponential*.
- We take into account the actual per iteration cost, and analyze not only the convergence rate of the algorithm, but also its numerical effort to solve a problem up to some relative precision. This allows us to classify the problems into categories where the optimal block size is one, the full matrix, or somewhere in between.
- We numerically validate the discovered theoretical properties of *determinantal sampling*, and demonstrate cases when it improves over uniform sampling, and when it performs similarly. If the two perform similarity, our analysis serves as a more interpretable proxy for the convergence analysis of uniform sampling.

1.2 Notation

Let S be a non-empty subset of $[d] := \{1, 2, \ldots, d\}$. We let $\mathbf{I}_{:S}$ be the $d \times |S|$ matrix composed of columns $i \in S$ of the $d \times d$ identity matrix $\mathbf{I}$. Note that $\mathbf{I}_{:S}^\top \mathbf{I}_{:S}$ is the $|S| \times |S|$ identity matrix. Given an invertible matrix $\mathbf{M} \in \mathbb{R}^{d\times d}$, we can extract its principal $|S| \times |S|$ sub-matrix with the corresponding rows and columns indexed by S via $\mathbf{M}_{SS} \overset{\text{def}}{=} \mathbf{I}_{:S}^\top \mathbf{M} \mathbf{I}_{:S}$, and additionally keeping the sub-matrix in its canonical place we can define the following operation,

$$\mathbf{M}_S \overset{\text{def}}{=} \mathbf{I}_{:S} \mathbf{M}_{SS} \mathbf{I}_{:S}^\top. \tag{1}$$

Note that $\mathbf{M}_S$ is the $n \times n$ matrix obtained from $\mathbf{M}$ by retaining elements $\mathbf{M}_{ij}$ for $i \in S$ and $j \in S$; and all the other elements set to zero. By $(\cdot)^+$ we denote the Moore-Penrose pseudoinverse. The matrix $(\mathbf{M}_S)^+$ can be calculated by inverting $\mathbf{M}_{SS} \in \mathbb{R}^{|S|\times|S|}$, and then placing it back into the $d \times d$ matrix.

2 ALGORITHM

The key assumption that motivates RNM is a smoothness condition that goes beyond the standard assumptions in the optimization literature, where smoothness would be characterized by a symmetric quadratic with the radius L. Instead, Assumption 1 below is tighter, allowing for more refined analysis, and can be related to the standard assumption by $L = \lambda_{\max}(\mathbf{M})$.

Assumption 1 (Smoothness)**.** *There exists a symmetric p.d. matrix* $\mathbf{M} \in \mathbb{R}^{d\times d}$ *such that* $\forall x, h \in \mathbb{R}^d$,

$$f(x+h) \leq f(x) + \langle \nabla f(x), h\rangle + \frac{1}{2}\langle h, \mathbf{M}h\rangle\,. \qquad (2)$$

This assumption is satisfied for quadratic problems such as ridge regression with squared loss, $y = \mathbf{A}^\top w + \epsilon$, where $\mathbf{A} \in \mathbb{R}^{n\times d}$ is the data matrix, and y is the vector of responses, which is corrupted via the noise $\epsilon \in \mathbb{R}^n$. In this case, Assumption 1 holds with $\mathbf{M}$ being the offset covariance matrix $\mathbf{A}^\top\mathbf{A} + \lambda\mathbf{I}$, where λ is the regularization parameter. Beyond quadratic problems, it holds for many common problems such as logistic regression, where $\mathbf{M} = \frac{1}{4}\mathbf{A}^\top\mathbf{A}$. Section 5.1 provides examples in the dual formulation.

2.1 Randomized Newton Method

Let k be the iteration count and x_0 be the initial point. The Randomized Newton Method algorithm is defined via the following update rule:

$$x_{k+1} = x_k - (\mathbf{M}_{S_k})^+ \nabla f(x_k), \qquad (3)$$

where $S_k \subseteq [d]$ is a subset of coordinates chosen at iteration k from random sampling $\hat{S}$ to be defined. Notice that since $\mathbf{M}_{S_k}$ is a sparse $d\times d$ matrix with only a $|S_k|\times|S_k|$ principal submatrix that is non-zero, its inversion costs $\mathcal{O}(|S_k|^3)$ arithmetic operations. Moreover, only $|S_k|$ elements of $\nabla f(x_k)$ are needed for the update. Note that if $|S_k| = 1$ then we are in the classical case of coordinate descent, while if $S = [d]$, then we are performing a Newton step (with $\mathbf{M}$ in place of the true Hessian).

2.2 Sampling

The strategy with which one chooses blocks $S_k \subseteq [d]$ in (3) is of great importance and it influences the algorithm significantly. This strategy, called a *sampling* and denoted $\hat{S}$, is a random set-valued mapping with values being subsets of $[d]$. A *proper sampling* is such that $p_i \overset{\text{def}}{=} \mathrm{P}(i \in \hat{S}) > 0$ for all i.

The most popular are *uniform* samplings, i.e., those for which the marginal probabilities are equal:

$$\mathrm{P}(i \in \hat{S}) = \mathrm{P}(j \in \hat{S}) \quad \forall i, j \in [d].$$

This class includes τ-nice and τ-list samplings (Qu and Richtárik, 2016). The τ-nice sampling considers all elements of a power set of $[d]$ with a fixed cardinality s.t. $|\hat{S}| = \tau$. There are $\binom{d}{\tau}$ of such subsets and each of them is equally probable. Consequently, the probability $\mathrm{P}(i \in \hat{S}) = \frac{\tau}{n}$. On the other hand, the τ-list sampling is restricted to ordered and consecutive subsets of the power set, with cardinality fixed to τ.

Data dependent (and potentially non-uniform) samplings, which sample according to the diagonal elements of $\mathbf{M}$, have been analyzed in the context of coordinate descent (Qu and Richtárik, 2016; Allen-Zhu et al., 2016; Hanzely and Richtárik, 2018; Richtárik and Takáč, 2015).

3 DETERMINANTAL SAMPLING

Our proposed sampling for the Randomized Newton Method is based on a class of distributions called *Determinantal Point Processes (DPPs)*. Originally proposed by Macchi (1975), DPPs have found numerous applications in machine learning (Kulesza and Taskar, 2012) as well as optimization (Zhang et al., 2017; Borsos et al., 2019), for their variance reduction properties and the ability to produce diverse samples.

Definition 1. *For a* $d\times d$ *p.s.d. matrix* $\mathbf{M}$*, we define* $\mathrm{DPP}(\mathbf{M})$ *as a distribution over all subsets* $S \subseteq [d]$*, so that*

$$\mathrm{P}(S) \propto \det(\mathbf{M}_{SS}). \qquad (4)$$

Even though this is a combinatorial distribution, the normalization constant can be computed exactly. We state this well known fact (e.g., see Kulesza and Taskar (2012)) separately because it is crucial for proving our main result.

Lemma 1 (Normalization)**.** *For a* $d\times d$ *matrix* $\mathbf{M}$*,*

$$\sum_{S\subseteq[d]} \det(\mathbf{M}_{SS}) = \det(\mathbf{I}+\mathbf{M}).$$

Note that the distribution samples out of a power set of $[d]$. While cardinality constrained versions have also been used, they lack certain properties such as a simple normalization constant. Even though the subset size of $\mathrm{DPP}(\mathbf{M})$ is a random variable, it is highly concentrated around its mean, and it can also be easily adjusted by replacing the matrix with a rescaled version $\frac{1}{\alpha}\mathbf{M}$, where $\alpha > 0$. This only affects the distribution of the subset sizes, with the expected size given by the following lemma (see Kulesza and Taskar, 2012).

Lemma 2 (Subset Size)**.** *If* $\hat{S} \sim \mathrm{DPP}\left(\frac{1}{\alpha}\mathbf{M}\right)$*, then*

$$\mathbb{E}[|\hat{S}|] = \mathrm{Trace}(\mathbf{M}(\alpha\mathbf{I}+\mathbf{M})^{-1}). \qquad (5)$$

By varying the value of α, we can obtain any desired expected subset size between 0 and d. As we increase α, the subset size decreases, whereas if we take $\alpha \to 0$, then in the limit the subset size becomes d, i.e., always selecting the $[d]$. While the relationship between α and $\mathbb{E}[|\hat{S}|]$ cannot be easily inverted analytically, it still provides a convenient way of smoothly interpolating between the full Newton and coordinate descent. To give a sense of what α can be used to ensure subset size bounded by some k, we give the following lemma.

Lemma 3. *Let $\{\lambda_i\}_{i=1}^d$ be the eigenvalues of $\mathbf{M}$ in a decreasing order. If $\alpha = \sum_{j \geq k} \lambda_j$, then $\mathbb{E}[|\hat{S}|] < k$.*

3.1 New expectation formula

We are now ready to state our main result regarding DPPs, which is a new expectation formula that can be viewed as a matrix counterpart of the determinantal identity from Lemma 1.

Theorem 1. *If $\mathbf{M} \succ \mathbf{0}$ and $\hat{S} \sim \mathrm{DPP}\left(\frac{1}{\alpha}\mathbf{M}\right)$, then*

$$\mathbb{E}\left[(\mathbf{M}_{\hat{S}})^{+}\right] = (\alpha\mathbf{I} + \mathbf{M})^{-1}. \tag{6}$$

Remark 1. *If we let $\mathbf{M} \succeq \mathbf{0}$, then the equality in (6) must be replaced by a p.s.d. inequality $\preceq$.*

We postpone the proof to the appendix. The remarkable simplicity of our result leads us to believe that it is of interest not only in the context of the Randomized Newton Method, but also to the broader DPP community. While some matrix expectation formulas involving the pseudoinverse have been recently shown for some special DPPs (e.g., Dereziński and Warmuth, 2018), this result for the first time relates an *unregularized* subsampled pseudoinverse with a $\alpha\mathbf{I}$-*regularized* inverse of the full matrix $\mathbf{M}$. Moreover, the amount of regularization that appears in the formula is directly related to the expected sample size.

3.2 Efficient sampling

Efficient DPP sampling has been an active area of research over the past decade. Several different approaches have been developed, such as an algorithm based on the eigendecomposition of $\mathbf{M}$ (Hough et al., 2006; Kulesza and Taskar, 2012) as well as an approximate MCMC sampler (Anari et al., 2016) among others. For our problem, it is important to be able to sample from $\mathrm{DPP}(\mathbf{M})$ without having to actually construct the entire matrix $\mathbf{M}$, and much faster than it takes to compute the full inverse $\mathbf{M}^{-1}$. Moreover, being able to rapidly generate multiple independent samples is crucial because of the iterative nature of the Randomized Newton Method. A recently proposed DPP sampler satisfies all of these conditions. We quote the time complexity of this method (the bounds hold with high probability relative to the randomness of the algorithm).

Lemma 4 (Dereziński et al. (2019))**.** *For a $d \times d$ p.s.d. matrix $\mathbf{M}$ let $k = \mathbb{E}[|\hat{S}|]$ where $\hat{S} \sim \mathrm{DPP}(\mathbf{M})$. Given $\mathbf{M}$, we can sample*

1. *the first $\hat{S}$ in: $d \cdot \mathrm{poly}(k)\,\mathrm{polylog}(d)$ time,*
2. *each next sample of $\hat{S}$ in: $\mathrm{poly}(k)$ time.*

Note that the time it takes to obtain the first sample (i.e., the preprocessing cost) is $o(d^2)$, meaning that we do not actually have to read the entire matrix $\mathbf{M}$. Moreover, the cost of producing repeated samples only depends on the sample size k, which is typically small. The key idea behind the algorithm of Dereziński et al. (2019) is to produce a larger sample of indices drawn i.i.d. proportionally to the marginal probabilities of $\mathrm{DPP}(\mathbf{M})$. For any $i \in [d]$, the marginal probability of i in $\hat{S} \sim \mathrm{DPP}(\frac{1}{\alpha}\mathbf{M})$ is:

$$\mathrm{P}(i \in \hat{S}) = \left[\mathbf{M}(\alpha\mathbf{I} + \mathbf{M})^{-1}\right]_{ii}.$$

In the randomized linear algebra literature, this quantity is often called the ith α-ridge leverage score (Alaoui and Mahoney, 2015), and sampling i.i.d. according to ridge leverage scores is known to have strong guarantees in approximating p.s.d. matrices.

Approximate ridge leverage score sampling incurs a smaller preprocessing cost compared to a DPP (Calandriello, 2017), and basically no resampling cost. Motivated by this, we propose to use this sampling as a fast approximation to $\mathrm{DPP}(\frac{1}{\alpha}\mathbf{M})$ and our experiments demonstrate that it exhibits similar convergence properties for Randomized Newton. We numerically compare the sampler from Lemma 4 against leverage score sampling in Appendix B.

4 CONVERGENCE ANALYSIS

In this section, we analyze the convergence properties of the update scheme (3) with determinantal sampling defined by (4). In order to establish linear rate of convergence, we need to assume strong convexity.

Assumption 2 (Strong Convexity)**.** *Under Assumption 1, there exists a $\kappa > 0$ such that $\forall x, h \in \mathbb{R}^d$,*

$$f(x) + \langle \nabla f(x), h\rangle + \frac{\kappa}{2}\langle h, \mathbf{M}h\rangle \leq f(x+h)$$

Intuitively, the parameter $\kappa \in (0, 1]$ measures the degree of accuracy of our quadratic approximation. For a quadratic function $\kappa = 1$.

Lemma 5 (Qu et al. (2016))**.** *Under Assumptions 1 and 2, let $\{x^k\}_{k \geq 0}$ be a sequence of random vectors produced by the Algorithm with a proper sampling $\hat{S}$,*

and let x^* be the optimum of f. Then,

$$\mathbb{E}\big[f(x^{k+1}) - f(x^*)\big] \leq \big(1 - \sigma(\hat{S})\big)\mathbb{E}\big[f(x^k) - f(x^*)\big],$$

where

$$\sigma(\hat{S}) \overset{def}{=} \kappa \cdot \lambda_{\min}\big(\mathbf{M}^{1/2}\mathbb{E}[(\mathbf{M}_{\hat{S}})^{+}]\mathbf{M}^{1/2}\big). \tag{7}$$

Strong convexity is not necessary to run RNM (3). In the cases where the function is only convex, we recover the standard sublinear rate depending on σ.

Lemma 6 (Karimireddy et al. (2018))**.** *Let f be convex and satisfy Assumption 1. Then using the update scheme in (3) with any proper sampling,*

$$\mathbb{E}[f(x^k) - f(x^*)] \leq \frac{2D}{\sigma(\hat{S})k}$$

where $\sigma(\hat{S})$ is as in (7), and $D = \max_x\{(x^* - x)^\top \mathbf{M}(x^* - x) | f(x) \leq f(x^0)\}$ is the set diameter in $\mathbf{M}$ geometry at the initial level sets.

The preceding two lemmas introduced the quantity $\sigma(\hat{S})$ characterizing the theoretical convergence rate of the method. By applying our new expectation formula (Theorem 1) we obtain a simple form for this quantity under DPP sampling.

Theorem 2. *Under Assumption 1, given $\alpha > 0$:*

$$\sigma(\hat{S}) = \kappa\frac{\lambda_d}{\lambda_d + \alpha} \quad \textit{for} \quad \hat{S} \sim \mathrm{DPP}\left(\tfrac{1}{\alpha}\mathbf{M}\right), \tag{8}$$

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

Note that $\sigma(\hat{S})$ depends solely on the smallest eigenvalue and the parameter α controlling the expected size. This is not the case for other samplings, and other closed forms are not known in general (Qu et al., 2016).

Recall that the smaller the α the bigger the subsets. The closed form expression from Theorem 2 combined with Lemma 3 allows us to formulate a recurrence relation between the convergence rates with different expected set sizes.

Proposition 1 (Recurrence relation)**.** *Let $\{\lambda_i\}_{i=1}^d$ be the eigenvalues of $\mathbf{M}$ in a decreasing order. Let $k < d$ be a positive integer, $\alpha(k) = \sum_{i>k-1}\lambda_i$, and $\sigma(k) = \frac{\lambda_d}{\lambda_d + \alpha(k)}$. Then,*

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

and $\sigma(d) = 1$ while $\mathbb{E}[|S|] < k$.

This result allows us to further improve the theoretical bounds from Qu et al. (2016) on the parameter σ. Namely, it has been previously established that σ grows at least linearly with the increasing subset size of τ-uniform sampling, i.e., $\tau\sigma(1) \leq \sigma(\tau)$. We can establish more informative bounds depending on the eigenvalue decay. Specifically, for a decreasing sequence of eigenvalues $\{\lambda_i\}_{i=1}^d$,

$$\left(1 + \sum_{j=1}^{\tau-1}\frac{\lambda_j}{\lambda_d}\right)\sigma(1) \leq \sigma(\tau). \tag{9}$$

For example, given exponentially decaying eigenvalues $\lambda_i = \gamma^i$ where $\gamma < 1$, the increase is at least exponential, and the convergence rate is at least $(1 + (\tau - 1)\gamma^{\tau-d})$ bigger. The case with linear speed-up is recovered when all eigenvalues are equal.

5 OPTIMAL BLOCK SIZE

Our results such as Proposition 1 and inequality (9) describe the convergence speedup of using larger coordinate blocks for RNM with determinantal sampling as a function of the eigenvalues of $\mathbf{M}$. In this section, we demonstrate that covariance matrices arising in kernel ridge regression have known asymptotic eigenvalue decays, which allows for a precise characterization of RNM performance.

5.1 Kernel Ridge Regression

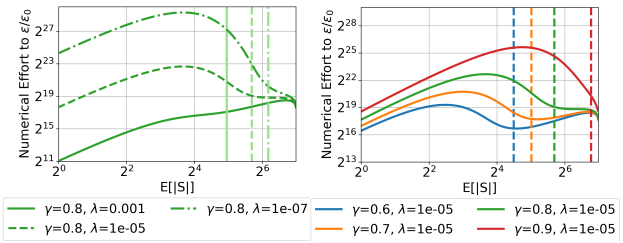
The motivating example for our analysis is the dual formulation of kernel ridge regression which is a natural application for block coordinate descent because of its high dimensionality. Suppose our (primal) regression problem is defined by the following objective:

$$\min_{x\in\mathbb{R}^d}\frac{1}{n}\sum_{i=1}^{n}\frac{1}{2}(\Phi(a_i)^\top x - y_i)^2 + \frac{\lambda}{2}\|x\|_2^2,$$

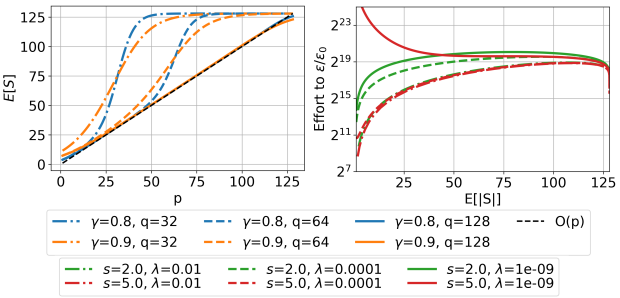
where $\Phi(\cdot)$ represents the kernel feature mapping and λ is the regularization parameter. Due to the Fenchel duality theorem (Borwein and Zhu, 2005), the dual formulation of this problem is:

$$\min_{\alpha\in\mathbb{R}^n}\frac{1}{2n}\alpha^\top\mathbf{K}\alpha + \frac{\lambda}{2}\sum_{i=1}^{n}\left(\alpha_i^2 + 2\alpha_i y_i\right), \tag{10}$$

where $\mathbf{K}_{ij} = \Phi(a_i)^\top\Phi(a_j)$. It is easy to see that the minimization problem (10) is exactly in the right form for RNM to be applied with the matrix $\mathbf{M} = \frac{1}{n}\mathbf{K} + \lambda\mathbf{I}$. Notice that $\mathbf{M}$ is an $n \times n$ matrix and sampling submatrices of $\mathbf{M}$ has the interpretation of subsampling the dataset. However, to keep the notation consistent with earlier discussion, w.l.o.g. we will let $d = n$ for the remainder of this section so that $\mathbf{M}$ is $d \times d$. We will also assume that the minimization problem is solved with the RNM update where each coordinate block is sampled as $\hat{S} \sim \mathrm{DPP}(\frac{1}{\alpha}\mathbf{M})$.



(a) The left column varies λ and the right one varies γ. The vertical line corresponds to the value of $q = \frac{\log(\lambda)}{\log(\gamma)}$.



(b) (left) $\mathbb{E}[|S|]$ vs parameter p for exponential kernel. (right) Numerical effort for polynomial decay.

Figure 1: In (a) we consider exponentially decreasing eigenvalues. In (b) (left) we plot the relationship between p and $\mathbb{E}[\hat{S}]$ for exponential decay. In (b) (right) we show the numerical effort for polynomial decay.

5.2 Exponentially decreasing spectrum

Let $\{\lambda_i\}_{i=1}^d$ be the eigenvalues of $\mathbf{M}$ in decreasing order. Suppose that the eigenvalue decay is exponentially decreasing:

$$\lambda_i = C\gamma^i + C\lambda \text{ for } \gamma < 1.$$

Motivation A classical motivating example for the exponential eigenvalue decay is the *squared exponential kernel*, where an analytical form of the decay can be derived for normally distributed data (Rasmussen and Williams, 2006). In particular, assuming $x \sim \mathcal{N}(0, \eta^2)$, and using the kernel function $k(x, y) = \exp\left(-\frac{(x-y)^2}{2l^2}\right)$ in one dimension, the eigenvalues satisfy $\lambda_k \leq C\gamma^k$ for a general constant C independent of k, where

$$\gamma = \frac{2\eta^2}{l^2 + 2\eta^2 + \sqrt{l^2 + 2\eta^2}}. \tag{11}$$

Complexity For the ease of exposition, suppose that $\gamma^{q+1} \leq \lambda \leq \gamma^q$, where λ is the regularization constant and $q \in [d]$. Intuitively, this means that the regularization parameter flattens the decay at γ^{q+1}, which will play a role in the analysis.

To control the expected size $\mathbb{E}[|\hat{S}|]$ of determinantal sampling, let $\alpha(p) = C\gamma^p$, where $p \in [1, d]$. We get:

$$\mathbb{E}[|\hat{S}|] \overset{(5)}{\leq} p + R_d(p, q) \quad \text{for} \quad \hat{S} \sim \text{DPP}\left(\tfrac{1}{\alpha(p)}\mathbf{M}\right),$$

where $R_d(p, q) = \sum_{i=1}^{d-p} \frac{\gamma^i + \gamma^{q-p}}{\gamma^i + \gamma^{q-p} + 1}$. Asymptotically, if $p \ll q$, i.e., the parameter $\alpha(p)$ dominates the regularization $C\lambda$, then the expected subset is $\mathbb{E}[|\hat{S}|] \approx p$. However, in the regime where $p = \Omega(q)$, the expected subset size rapidly goes up to d (see Figure 1b (left)).

We now derive the convergence rate of RNM under determinantal sampling:

$$1 - \sigma(\hat{S}) \overset{(8)}{=} \frac{C\gamma^p}{C\gamma^d + C\lambda + C\gamma^p} \geq \frac{1}{1 + \gamma^{-p}(\gamma^d + \gamma^q)}.$$

Likewise one can see that the convergence rate improves exponentially with p.

Numerical effort From Theorem 5, we know that in order to reach an ϵ accurate solution from the initial accuracy $\epsilon_0 = f(x^0) - f(x^*)$ under the convergence rate $\epsilon \leq (1 - \sigma)^T \epsilon_0$, the number of needed steps can be bounded by

$$T \leq \frac{\log(\epsilon_t) - \log(\epsilon_0)}{\log(1 - \sigma)}. \tag{12}$$

Using the bound derived for $1 - \sigma(\hat{S})$, we obtain $T \leq \left(\log(\frac{1}{\epsilon_t}) - \log(\frac{1}{\epsilon_0})\right)\left(\frac{\gamma^p}{\gamma^d + \gamma^q} + 1\right)$. Since, the computation step is dominated by the inversion operation $\mathcal{O}(\mathbb{E}[|\hat{S}|]^3)$, the number of arithmetic operations is

$$\mathcal{O}\left(\mathbb{E}[|S|]^3 \cdot T\right) \leq \mathcal{O}\left(\left(p + R_d(p, q)\right)^3 \frac{\gamma^p}{\gamma^d + \gamma^q}\right).$$

The upper bound on the numerical effort in the previous equation has two regimes. At first, for small subset sizes it is increasing, but then exponential decay starts to dominate and using larger blocks significantly improves the convergence rate. Finally it flattens around $\mathbb{E}[|\hat{S}|] = q$. Note that when $\lambda \approx \gamma^q$, i.e., for $q = \frac{\log(\lambda)}{\log(\gamma)}$, this phenomenon is visualized in Figure 1a where the vertical bars correspond to q. In the regime where $d \approx q$, inverting the whole matrix seems to be the best option. When $q < d$, the term γ^q dominates the term γ^d, and the best subset size is either 1 or on the order of q, depending on the value of λ.

These observations are contrary to the intuition from the previous works. We suspect that, due to fixed memory fetching costs, for small sizes the initial phase is unobserved but the second phase should be observed. Figure 1a suggests that for sufficiently small values of λ the numerical performance is maximized at the attenuation point q and the predicted optimal block size is $\frac{\log(\lambda)}{\log(\gamma)}$.

5.3 Polynomially decaying spectrum

Suppose that the eigenvalues $\{\lambda_i\}_{i=1}^d$ are decreasing polynomially, i.e., so that $\lambda_i = Ci^{-s} + C\lambda$ for $s > 1$.

Motivation For example, consider a Matérn kernel of order s, which has the form $k(x,y) = C_2 B_s(\|x-y\|)\exp(-C_1\|x-y\|)$, where C_1, C_2 are constants, and $B_s(d)$ is a modified Bessel function of order s (see Rasmussen and Williams, 2006). This class of kernels exhibits asymptotically polynomial decay of eigenvalues (see Seeger et al., 2008).

Complexity Suppose, $\lambda = q^{-s}$ for $q \in [1,\infty)$. To control the expected size let us parameterize the tuning parameters as $\alpha = Cp^{-s}$, where C is a suitable general constant. Then the convergence rate becomes:

$$1 - \sigma(\hat{S}) = \frac{p^{-s}}{d^{-s} + p^{-s} + q^{-s}} = \frac{1}{(p/d)^s + (p/q)^s + 1}$$

and $\mathbb{E}[|\hat{S}|] = \sum_{i=1}^d \frac{i^{-s}+q^{-s}}{i^{-s}+p^{-s}+q^{-s}}$. If $p \ll q$, we can establish by integral approximation that $\mathbb{E}[|S|] \approx \mathcal{O}(p)$, otherwise the expected size grows faster. Additionally, with increasing p the convergence rate always improves.

Numerical Effort When $p \ll q$, similarly as in the preceding subsection, the numerical cost becomes $\mathcal{O}\big(p^3(\frac{d^s q^s}{p^s(q^s+d^s)})\big)$. This suggests that for $s \geq 3$ the total numerical cost decreases for larger subsets, while for the problems with smaller s, the cost increases. In general, it is difficult to obtain general insights from the formulas, but the visalization in Figure 1b (right) suggests that if the regularization constant is large (small q), even problems with large s might incur more cost as the subset size increases.

This suggests that *small block sizes matching the memory fetching costs* should be optimal if either the regularization is large or if s is small. With the same assumption, if the desired accuracy is very high, performing *full matrix inversion* can be more efficient, corresponding to $\mathbb{E}[|\hat{S}|] \to d$ in Figure 1b (right). Note that increasing the accuracy to which we optimize the problem shifts the curves up in the logarithmic plot, while keeping the end point fixed.

5.4 Sparse spectrum

Suppose that only s out of the d eigenvalues are relatively large, while the remaining ones are very small. This scenario occurs with a linear kernel where the number of large eigenvalues corresponds to the number of features, and the remaining ones are proportional to the regularization parameter λ.

Complexity For the ease of exposition, let the large eigenvalues all be equal to $\mu \gg \lambda$. Lemma 5 implies that if $\alpha = \sum_{i>k-1}^d \lambda_i$ then $\mathbb{E}[|\hat{S}|] \leq k$. The convergence rate can be split to two cases:

$$1-\sigma(\hat{S}) = \begin{cases} \frac{d-k}{d-k+1} & \text{when } k \in [s, d-1], \\ 1 - \frac{\lambda}{\mu}\frac{1}{d-k} + \mathcal{O}\left(\frac{\lambda^2}{\mu^2}\right) & \text{when } k \in [0, s). \end{cases}$$

We see that once $k \geq s$ a discontinuity in the spectrum implies a discontinuity in the convergence rate. Consequently, the *optimal subset size is of the order of* s as long as $\frac{\lambda}{\mu}$ is sufficiently small.

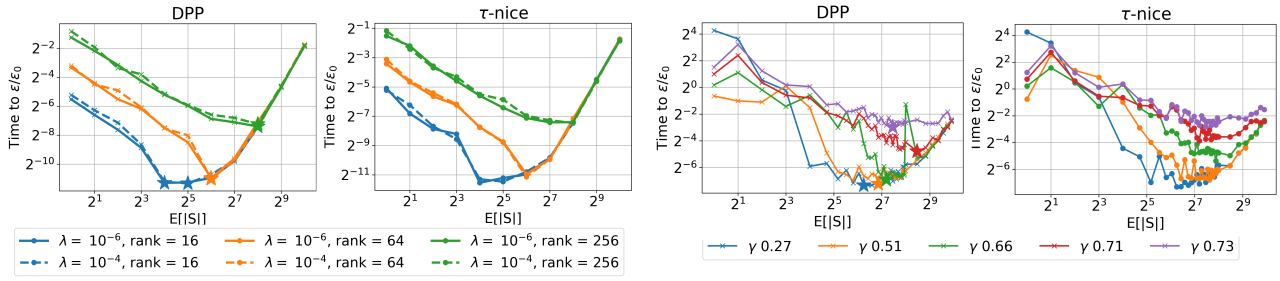
6 EXPERIMENTS

We numerically validate the theoretical findings from the previous sections. Our main objective is to demonstrate that the convergence behavior of RNM under DPP sampling aligns well with the behavior of RNM under uniform sampling (called τ-nice), which is more commonly used. This would suggest that our convergence analysis under DPP sampling is also predictive for other standard samplings. In addition to providing evidence for this claim, we also show that there are cases where DPP sampling leads to superior performance of RNM.

Even though there exist efficient algorithms for DPP sampling, we chose to use approximate ridge leverage score sampling as a cheaper proxy for DPP sampling, as suggested in a recent line of work (Dereziński, 2019; Dereziński et al., 2019). The real data experiments were performed with sampling according to the $\frac{1}{2}$-approximate ridge leverage scores (Calandriello, 2017). We always report the mean value of 10 reruns of the experiment with the given parameters.

Gaussian Data The first experiment deals with data sampled from a Gaussian distribution. The optimization using a kernel $\mathbf{K}$ with sparse spectrum (Figure 2a) verifies the theoretical findings that the optimal block size should be of the same order as rank($\mathbf{K}$). Using similarly generated data, and the relation in (11) to relate lengthscale l and γ of squared exponential kernel, we reproduce the prediction of the theory that for sharper decays the optimal expected size should be larger (see Figure 2b, compared with theory, Figure 1a). The performance of DPP and uniform sampling is on par as the intuition suggests, since for normally distributed data even a uniform subsample provides good summary statistics.

Gaussian Mixture Data Akin to results from the sketching literature (e.g., see Dereziński and Warmuth, 2018), we suspect that the superior convergence of DPP sampling over uniform presents itself primarily if the



(a) Sparse spectrum, rank of $\mathbf{K}$ shown.

(b) Exponential decay, varying lengthscale γ, for $\lambda = 10^{-7}$.

Figure 2: For Gaussian data, RNM exhibits similar behavior under DPP and uniform (τ-nice) samplings.

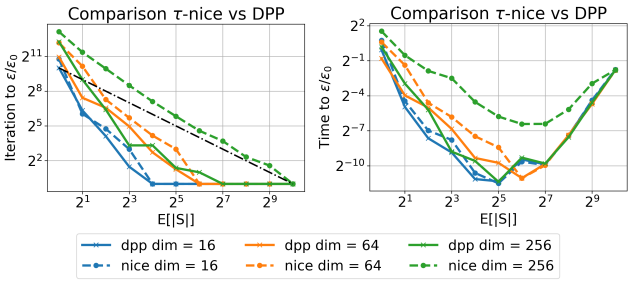
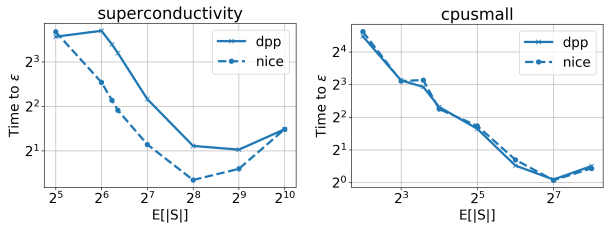


Figure 3: Gaussian mixture data - sparse spectrum, rank of kernel $\mathbf{K}$ shown.



(a) $n = 1460$ and $d = 81$.

(b) $n = 2191$ and $d = 12$.

Figure 4: Experiments on real data.

dataset is heterogeneous. By heterogeneity we mean that a uniform subsampling of the points is likely not a good summary of the dataset. Consider a dataset where the points are sampled from a Gaussian Mixture Model with 8 clusters that are equally likely. In order to have a good summary, a point from each cluster should be present in the sample. DPP samples are generally more diverse than uniform samples which makes it more likely that they will cover all the clusters. In Figure 3, we see that DPP significantly outperforms uniform sampling for this dataset because it allows RNM to solve more representative subproblems.

Real Data Experiments We perform two real data experiments on standard UCI datasets where we optimize until statistical precision. In Figure 4a, we optimize linear ridge regression on the *superconductivity* dataset. Next, in Figure 4b we fit kernel ridge regression with squared exponential kernel on the *cpusmall* dataset. For both datasets, the optimal subset size under DPP sampling roughly matches the optimal size under uniform sampling. Moreover, in the case of the superconductivity dataset, as suggested by the theory for linear kernels, the optimal size is of the same order as the feature dimensionality.

7 CONCLUSION

We analyzed a sampling strategy for the Randomized Newton Method, where coordinate blocks of the Hessian over-approximation are sampled according to their determinant. This sampling allows for a simple interpretation of the convergence rate of the algorithm, which was previously not well understood. We demonstrated that for empirical risk minimization this convergence analysis allows us to predict the optimal size for the sampled coordinate blocks in order to minimize the total computational cost of the optimization.

Acknowledgments This work was supported by SNSF grant 407540_167212 through the NRP 75 Big Data program. Also, MD thanks the NSF for funding via the NSF TRIPODS program.

References

Alaoui, A. E. and Mahoney, M. W. (2015). Fast randomized kernel ridge regression with statistical guarantees. In *Proceedings of the 28th International Conference on Neural Information Processing Systems*, pages 775–783, Montreal, Canada.

Allen-Zhu, Z., Qu, Z., Richtárik, P., and Yuan, Y.

(2016). Even faster accelerated coordinate descent using non-uniform sampling. In *International Conference on Machine Learning*, pages 1110–1119.

Anari, N., Gharan, S. O., and Rezaei, A. (2016). Monte carlo markov chain algorithms for sampling strongly rayleigh distributions and determinantal point processes. In Feldman, V., Rakhlin, A., and Shamir, O., editors, *29th Annual Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pages 103–115, Columbia University, New York, New York, USA. PMLR.

Blanchard, G., Bousquet, O., and Zwald, L. (2007). Statistical properties of kernel principal component analysis. *Machine Learning*, 66(2):259–294.

Borsos, Z., Curi, S., Levy, K. Y., and Krause, A. (2019). Online variance reduction with mixtures. In Chaudhuri, K. and Salakhutdinov, R., editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 705–714, Long Beach, California, USA. PMLR.

Borwein, J. and Zhu, Q. (2005). *Techniques of Variational Analysis*. Springer-Verlag New York.

Calandriello, D. (2017). *Efficient Sequential Learning in Structured and Constrained Environments*. PhD thesis, Inria Lille Nord Europe-Laboratoire CRIStAL-Université de Lille.

Dereziński, M. (2019). Fast determinantal point processes via distortion-free intermediate sampling. In Beygelzimer, A. and Hsu, D., editors, *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 1029–1049, Phoenix, USA.

Dereziński, M., Calandriello, D., and Valko, M. (2019). Exact sampling of determinantal point processes with sublinear time preprocessing. In Wallach, H., Larochelle, H., Beygelzimer, A., d Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems 32*, pages 11542–11554. Curran Associates, Inc.

Dereziński, M. and Warmuth, M. K. (2018). Reverse iterative volume sampling for linear regression. *Journal of Machine Learning Research*, 19(23):1–39.

Dünner, C., Forte, S., Takáč, M., and Jaggi, M. (2016). Primal-dual rates and certificates. In *Proceedings of The 33rd International Conference on Machine Learning*, pages 783–792.

Gower, R., Koralev, D., Lieder, F., and Richtarik, P. (2019). RSN: Randomized subspace Newton. In Wallach, H., Larochelle, H., Beygelzimer, A., d Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems 32*, pages 614–623. Curran Associates, Inc.

Gower, R. M. and Richtárik, P. (2015). Randomized iterative methods for linear systems. *SIAM. J. Matrix Anal. & Appl., 36(4), 1660–1690, 2015.*

Hanzely, F. and Richtárik, P. (2018). Accelerated coordinate descent with arbitrary sampling and best rates for minibatches. *arXiv preprint arXiv:1809.09354.*

Hough, J. B., Krishnapur, M., Peres, Y., Virág, B., et al. (2006). Determinantal processes and independence. *Probability surveys*, 3:206–229.

Johnson, R. and Zhang, T. (2013). Accelerating stochastic gradient descent using predictive variance reduction. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'13, pages 315–323, USA. Curran Associates Inc.

Karimireddy, S. P. R., Stich, S., and Jaggi, M. (2018). Adaptive balancing of gradient and update computation times using global geometry and approximate subproblems. In Storkey, A. and Perez-Cruz, F., editors, *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pages 1204–1213, Playa Blanca, Lanzarote, Canary Islands. PMLR.

Kulesza, A. and Taskar, B. (2012). *Determinantal Point Processes for Machine Learning*. Now Publishers Inc., Hanover, MA, USA.

Ma, C., Smith, V., Jaggi, M., Jordan, M., Richtarik, P., and Takac, M. (2015). Adding vs. averaging in distributed primal-dual optimization. In Bach, F. and Blei, D., editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1973–1982, Lille, France. PMLR.

Macchi, O. (1975). The coincidence approach to stochastic point processes. *Advances in Applied Probability*, 7(1):83–122.

Mutný, M. and Richtárik, P. (2018). Parallel stochastic newton method. *Journal of Computational Mathematics*, 36(3):405–426.

Nguyen, L. M., Liu, J., Scheinberg, K., and Takáč, M. (2017). Sarah: A novel method for machine learning problems using stochastic recursive gradient. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2613–2621. JMLR. org.

Pilanci, M. and Wainwright, M. J. (2015). Newton sketch: A linear-time optimization algorithm with linear-quadratic convergence. *arXiv preprint arXiv:1505.02250*.

Qu, Z., Richtárik, P., Takáč, M., and Fercoq, O. (2016). SDNA: Stochastic Dual Newton Ascent for Empirical Risk Minimization. *Proceedings of The 33rd International Conference on Machine Learning*.

Qu, Z. and Richtárik, P. (2016). Coordinate descent with arbitrary sampling ii: Expected separable overapproximation. *Optimization Methods and Software*, 31(5):858–884.

Rasmussen, C. and Williams, C. (2006). Gaussian processes for machine learning. *The MIT Press, Cambridge, doi*, 10:S0129065704001899.

Richtárik, P. and Takáč, M. (2015). On optimal probabilities in stochastic coordinate descent methods. *Optimization Letters*, pages 1–11.

Richtárik, P. and Takáč, M. (2016). Parallel coordinate descent methods for big data optimization. *Mathematical Programming*, 156(1):433–484.

Richtárik, P. and Takáč, M. (2016). Distributed coordinate descent method for learning with big data. *J. Mach. Learn. Res.*, 17(1):2657–2681.

Rodomanov, A. and Kropotov, D. (2019). A randomized coordinate descent method with volume sampling. *arXiv preprint arXiv:1904.04587*.

Roosta-Khorasani, F. and Mahoney, M. W. (2016). Sub-Sampled Newton Methods I: Globally Convergent Algorithms. *ArXiv e-prints*.

Roosta-Khorasani, F. and Mahoney, M. W. (2016). Sub-sampled newton methods ii: Local convergence rates. *arXiv preprint arXiv:1601.04738*.

Seeger, M. W., Kakade, S. M., and Foster, D. P. (2008). Information consistency of nonparametric gaussian process methods. *IEEE Transactions on Information Theory*, 54(5):2376–2382.

Shalev-Shwartz, S. and Ben-David, S. (2014). *Understanding machine learning: From theory to algorithms*. Cambridge university press.

Shalev-Shwartz, S. and Zhang, T. (2013). Stochastic dual coordinate ascent methods for regularized loss. *The Journal of Machine Learning Research*, 14(1):567–599.

Zhang, C., Kjellström, H., and Mandt, S. (2017). Determinantal point processes for mini-batch diversification. In *33rd Conference on Uncertainty in Artificial Intelligence, UAI 2017, Sydney, Australia, 11 August 2017 through 15 August 2017*. AUAI Press Corvallis.