Structured adaptive and random spinners for fast machine learning computations (Supplementary Material)

In the Supplementary material we prove all theorems presented in the main body of the paper.

5.4 Structured machine learning algorithms with *Structured Spinners*

We prove now Lemma 1, Remark 1, as well as Theorem 1 and Theorem 3.

5.4.1 Proof of Remark 1

This result first appeared in [Ailon and Chazelle, 2006]. The following proof was given in [Choromanski and Sindhwani, 2016], we repeat it here for completeness. We will use the following standard concentration result.

Lemma 2 *(Azuma's Inequality) Let $X_1, ..., X_n$ be a martingale and assume that $-\alpha_i \leq X_i \leq \beta_i$ for some positive constants $\alpha_1, ..., \alpha_n, \beta_1, ..., \beta_n$. Denote $X = \sum_{i=1}^{n} X_i$. Then the following is true:*

$$\mathbb{P}[|X - \mathbb{E}[X]| > a] \leq 2e^{-\frac{a^2}{2\sum_{i=1}^{n}(\alpha_i+\beta_i)^2}} \quad (4)$$

Proof: Denote by $\tilde{\mathbf{x}}^j$ an image of $\mathbf{x}^j$ under transformation $\mathbf{HD}$. Note that the i^{th} dimension of $\tilde{\mathbf{x}}^j$ is given by the formula: $\tilde{x}_i^j = h_{i,1}x_1^j + ... + h_{i,n}x^{j,n}$, where $h_{l,u}$ stands for the l^{th} element of the u^{th} column of the randomized Hadamard matrix $\mathbf{HD}$. First, we use Azuma's Inequality to find an upper bound on the probability that $|\tilde{x}_i^j| > a$, where $a = \frac{\log(n)}{\sqrt{n}}$. By Azuma's Inequality, we have:

$$\mathbb{P}[|h_{i,1}x_1^j + ... + h_{i,n}x^{j,n}| \geq a] \leq 2e^{-\frac{\log^2(n)}{8}}. \quad (5)$$

We use: $\alpha_i = \beta_i = \frac{1}{\sqrt{n}}$. Now we take the union bound over all n dimensions and the proof is completed. □

5.4.2 *Structured Spinners*-equivalent definition

We will introduce here an equivalent definition of the model of structured spinners that is more technical (thus we did not give it in the main body of the paper), yet more convenient to work with in the proofs.

Note that from the definition of structured spinners we can conclude that each structured matrix $\mathbf{G}_{struct} \in \mathbb{R}^{n\times n}$ from the family of structured spinners is a product of three main structured blocks, i.e.:

$$\mathbf{G}_{struct} = \mathbf{B}_3\mathbf{B}_2\mathbf{B}_1, \quad (6)$$

where matrices $\mathbf{B}_1, \mathbf{B}_2, \mathbf{B}_3$ satisfy two conditions that we give below.

Condition 1: Matrices: $\mathbf{B}_1$ and $\mathbf{B}_2\mathbf{B}_1$ are $(\delta(n), p(n))$-balanced isometries.
Condition 2: Pair of matrices $(\mathbf{B}_2, \mathbf{B}_3)$ is $(K, \Lambda_F, \Lambda_2)$-random.

Below we give the definition of $(K, \Lambda_F, \Lambda_2)$-randomness.

Definition 4 ($(K, \Lambda_F, \Lambda_2)$-randomness) *A pair of matrices $(\boldsymbol{Y}, \boldsymbol{Z}) \in \mathbb{R}^{n\times n} \times \mathbb{R}^{n\times n}$ is $(K, \Lambda_F, \Lambda_2)$-random if there exists $\boldsymbol{r} \in \mathbb{R}^k$, and a set of linear isometries $\phi = \{\phi_1, ..., \phi_n\}$, where $\phi_i : \mathbb{R}^n \to \mathbb{R}^k$, such that:*

- *$\boldsymbol{r}$ is either a ± 1-vector with i.i.d. entries or Gaussian with identity covariance matrix,*
- *for every $\boldsymbol{x} \in \mathbb{R}^n$ the j^{th} element $(\boldsymbol{Zx})_j$ of $\boldsymbol{Zx}$ is of the form: $\boldsymbol{r}^T \cdot \phi_j(\boldsymbol{x})$,*
- *there exists a set of i.i.d. sub-Gaussian random variables $\{\rho_1, ..., \rho_n\}$ with sub-Gaussian norm at most K, mean 0, the same second moments and a (Λ_F, Λ_2)-smooth set of matrices $\{\boldsymbol{W}^i\}_{i=1,...,n}$ such that for every $\boldsymbol{x} = (x_1, ..., x_n)^T$, we have: $\phi_i(\boldsymbol{Yx}) = \boldsymbol{W}^i(\rho_1 x_1, ..., \rho_n x_n)^T$.*

5.4.3 Proof of Lemma 1

Proof: Let us first assume the $\mathbf{G}_{circ}\mathbf{D}_2\mathbf{HD}_1$-setting (analysis for Toeplitz Gaussian or Hankel Gaussian is completely analogous). In that setting, it is easy to see that one can take $\mathbf{r}$ to be a Gaussian vector (this vector corresponds to the first row of $\mathbf{G}_{circ}$). Furthermore linear mappings ϕ_i are defined as: $\phi_i((x_0, x_1, ..., x_{n-1})^T) = (x_{n-i}, x_{n-i+1}, ..., x_{i-1})^T$, where operations on indices are modulo n. The value of $\delta(n)$ and $p(n)$ come from the fact that matrix $\mathbf{HD}_1$ is used as a $(\delta(n), p(n))$-balanced matrix and from Remark 1. In that setting, sequence $(\rho_1, ..., \rho_n)$ is discrete and corresponds to the diagonal of $\mathbf{D}_2$. Thus we have: $K = 1$. To calculate Λ_F and Λ_2, note first that matrix $\mathbf{W}^1$ is defined as $\mathbf{I}$ and subsequent $\mathbf{W}^i$s are given as circulant shifts of the previous ones (i.e. each row is a circulant shift of the previous row). That observation comes directly from the circulant structure of $\mathbf{G}_{circ}$. Thus we have: $\Lambda_F = O(\sqrt{n})$ and $\Lambda_2 = O(1)$. The former is true since each $\mathbf{A}^{i,j}$ has $O(n)$ nonzero entries and these are all 1s. The latter is true since each

nontrivial $\mathbf{A}^{i,j}$ in that setting is an isometry (this is straightforward from the definition of $\{\mathbf{W}^i\}_{i=1,...,n}$). Finally, all other conditions regarding $\mathbf{W}^i$-matrices are clearly satisfied (each column of each $\mathbf{W}^i$ has unit L_2 norm and corresponding columns from different $\mathbf{W}^i$ and $\mathbf{W}^j$ are clearly orthogonal).

Now let us consider the setting, where the structured matrix is of the form: $\sqrt{n}\mathbf{HD}_3\mathbf{HD}_2\mathbf{HD}_1$. In that case, $\mathbf{r}$ corresponds to a discrete vector (namely, the diagonal of $\mathbf{D}_3$). Linear mappings ϕ_i are defined as: $\phi_i((x_1,...,x_n)^T) = (\sqrt{n}h_{i,1}x_1,...,\sqrt{n}h_{i,n}x_n)^T$, where $(h_{i,1},...,h_{i,n})^T$ is the i^{th} row of $\mathbf{H}$. One can also notice that the set $\{\mathbf{W}^i\}_{i=1,...,n}$ is defined as: $w^i_{a,b} = \sqrt{n}h_{i,a}h_{a,b}$. Let us first compute the Frobenius norm of the matrix $\mathbf{A}^{i,j}$, defined based on the aforementioned sequence $\{\mathbf{W}^i\}_{i=1,...,n}$. We have:

$$\|\mathbf{A}^{i,j}\|_F^2 = \sum_{l,t\in\{1,...,n\}} (\sum_{k=1}^{n} w^j_{k,l}w^i_{k,t})^2$$
$$= n^2 \sum_{l,t\in\{1,...,n\}} (\sum_{k=1}^{n} h_{j,k}h_{k,l}h_{i,k}h_{k,t})^2 \quad (7)$$

To compute the expression above, note first that for $r_1 \neq r_2$ we have:

$$\theta = \sum_{k,l} h_{r_1,k}h_{r_1,l}h_{r_2,k}h_{r_2,l}$$
$$= \sum_{k} h_{r_1,k}h_{r_2,k} \sum_{l} h_{r_1,l}h_{r_2,l} = 0, \quad (8)$$

where the last equality comes from fact that different rows of H are orthogonal. From the fact that $\theta = 0$ we get:

$$\|\mathbf{A}^{i,j}\|_F^2 = n^2 \sum_{r=1,...,n} \sum_{k,l} h^2_{i,r}h^2_{j,r}h^2_{r,k}h^2_{r,l}$$
$$= n \cdot n^2 (\frac{1}{\sqrt{n}})^8 \cdot n^2 = n. \quad (9)$$

Thus we have: $\Lambda_F \leq \sqrt{n}$.

Now we compute $\|\mathbf{A}^{i,j}\|_2$. Notice that from the definition of $\mathbf{A}^{i,j}$ we get that

$$\mathbf{A}^{i,j} = \mathbf{E}^{i,j}\mathbf{F}^{i,j}, \quad (10)$$

where the l^{th} row of $\mathbf{E}^{i,j}$ is of the form $(h_{j,1}h_{1,l},...,h_{j,n}h_{n,l})$ and the t^{th} column of $\mathbf{F}^{i,j}$ is of the form $(h_{i,1}h_{1,t},...,h_{i,n}h_{n,t})^T$. Thus one can easily verify that $\mathbf{E}^{i,j}$ and $\mathbf{H}^{i,j}$ are isometries (since $\mathbf{H}$ is) thus $\mathbf{A}^{i,j}$ is also an isometry and therefore $\Lambda_2 = 1$. As in the previous setting, remaining conditions regarding matrices $\mathbf{W}^i$ are trivially satisfied (from the basic properties of Hadamard matrices). That completes the proof. □

5.4.4 Proof of Theorem 1

Let us briefly give an overview of the proof before presenting it in detail. Challenges regarding proving accuracy results for structured matrices come from the fact that, for any given $\mathbf{x} \in \mathbb{R}^n$, different dimensions of $\mathbf{y} = \mathbf{G}_{struct}\mathbf{x}$ are no longer independent (as it is the case for the unstructured setting). For matrices from the family of structured spinners we can, however, show that with high probability different elements of $\mathbf{y}$ correspond to projections of a given vector $\mathbf{r}$ (see Section 3) into directions that are close to orthogonal. The "close-to-orthogonality" characteristic is obtained with the use of the Hanson-Wright inequality that focuses on concentration results regarding quadratic forms involving vectors of sub-Gaussian random variables. If $\mathbf{r}$ is Gaussian, then from the well-known fact that projections of the Gaussian vector into orthogonal directions are independent, we can conclude that dimensions of $\mathbf{y}$ are "close to independent". If $\mathbf{r}$ is a discrete vector then we need to show that for n large enough, it "resembles" the Gaussian vector. This is where we need to apply the aforementioned techniques regarding multivariate Berry-Esseen-type central limit theorem results.

Proof: We will use notation from Section 3 and previous sections of the Supplement. We assume that the model with structured matrices stacked vertically, each of m rows, is applied. Without loss of generality, we can assume that we have just one block since different blocks are chosen independently. Let $\mathbf{G}_{struct}$ be a matrix from the family of structured spinners. Let us assume that $\mathbf{G}_{struct}$ is used by a function f operating in the d-dimensional space and let us denote by $\mathbf{x}^1,\ldots,\mathbf{x}^d$ some fixed orthonormal basis of that space. Our first goal is to compute: $\mathbf{y}^1 = \mathbf{G}_{struct}\mathbf{x}^1,...,\mathbf{y}^d = \mathbf{G}_{struct}\mathbf{x}^d$. Denote by $\tilde{\mathbf{x}}^i$ the linearly transformed version of $\mathbf{x}$ after applying block $\mathbf{B}_1$, i.e. $\tilde{\mathbf{x}}^i = \mathbf{B}_1\mathbf{x}^i$. Since $\mathbf{B}_1$ is $(\delta(n),p(n))$-balanced), we conclude that with probability at least: $p_{balanced} \geq 1 - dp(n)$ each element of each $\tilde{\mathbf{x}}^i$ has absolute value at most $\frac{\delta(n)}{\sqrt{n}}$. We shortly say that each $\tilde{\mathbf{x}}^i$ is $\delta(n)$-balanced. We call this event $\mathcal{E}_{balanced}$.

Note that by the definition of structured spinners, each $\mathbf{y}^i$ is of the form:

$$\mathbf{y}^i = (\mathbf{r}^T \cdot \phi_1(\mathbf{B}_2\tilde{\mathbf{x}}^i),...,\mathbf{r}^T \cdot \phi_m(\mathbf{B}_2\tilde{\mathbf{x}}^i))^T. \quad (11)$$

For clarity and to reduce notation, we will assume that $\mathbf{r}$ is n-dimensional. To obtain results for vectors $\mathbf{r}$ of different dimensionality D, it suffices to replace in our analysis and theoretical statements n by D. Let us denote $\mathcal{A} = \{\phi_1(\mathbf{B}_2\tilde{\mathbf{x}}^1),...,\phi_m(\mathbf{B}_2\tilde{\mathbf{x}}^1),...,\phi_1(\mathbf{B}_2\tilde{\mathbf{x}}^d),...,\phi_m(\mathbf{B}_2\tilde{\mathbf{x}}^d))\}$. Our goal is to show that with high probability (in respect to random choices of $\mathbf{B}_1$ and $\mathbf{B}_2$) for all $\mathbf{v}^i,\mathbf{v}^j \in \mathcal{A}$, $i \neq j$ the following is true:

$$|(\mathbf{v}^i)^T \cdot \mathbf{v}^j| \leq t \quad (12)$$

for some given $0 < t \ll 1$.

Fix some $t > 0$. We would like to compute the lower bound on the corresponding probability. Let us fix two vectors $\mathbf{v}^1, \mathbf{v}^2 \in \mathcal{A}$ and denote them as: $\mathbf{v}^1 = \phi_i(\mathbf{B}_2\mathbf{x})$, $\mathbf{v}^2 = \phi_j(\mathbf{B}_2\mathbf{y})$ for some $\mathbf{x} = (x_1, ..., x_n)^T$ and $\mathbf{y} = (y_1, ..., y_n)^T$. Note that we have (see denotation from Section 3):

$$\phi_i(\mathbf{B}_2\mathbf{x}) = (w^i_{11}\rho_1 x_1 + ... + w^i_{1,n}\rho_n x_n, ..., w^i_{n,1}\rho_1 x_1 + ... + w^i_{n,n}\rho_n x_n)^T \quad (13)$$

and

$$\phi_j(\mathbf{B}_2\mathbf{y}) = (w^j_{11}\rho_1 y_1 + ... + w^j_{1,n}\rho_n y_n, ..., w^j_{n,1}\rho_1 y_1 + ... + w^j_{n,n}\rho_n y_n)^T. \quad (14)$$

We obtain:

$$(\mathbf{v}^1)^T \cdot \mathbf{v}^2 = \sum_{l\in\{1,...,n\}, u\in\{1,...,n\}} \rho_l\rho_u(\sum_{k=1}^{n} x_l y_u w^i_{k,u} w^j_{k,l}). \quad (15)$$

We now show that, under assumptions from Theorem 1, the expected value of the expression above is 0. We have:

$$\mathbb{E}[(\mathbf{v}^1)^T \cdot \mathbf{v}^2] = \mathbb{E}[\sum_{l\in\{1,...,n\}} \rho_l^2 x_l y_l (\sum_{k=1}^{n} w^i_{k,l} w^j_{k,l})], \quad (16)$$

since $\rho_1, ..., \rho_n$ are independent and have expectations equal to 0. Now notice that if $i \neq j$ then from the assumption that corresponding columns of matrices $\mathbf{W}^i$ and $\mathbf{W}^j$ are orthogonal, we get that the above expectation is 0. Now assume that $i = j$. But then $\mathbf{x}$ and $\mathbf{y}$ have to be different and thus they are orthogonal (since they are taken from the orthonormal system transformed by an isometry). In that setting we get:

$$\mathbb{E}[(\mathbf{v}^1)^T \cdot \mathbf{v}^2] = \mathbb{E}[\sum_{l\in\{1,...,n\}} \rho_l^2 x_l y_l (\sum_{k=1}^{n} (w^i_{k,l})^2)] = \tau w \sum_{l=1}^{n} x_l y_l = 0, \quad (17)$$

where τ stands for the second moment of each ρ_i, w is the squared L_2-norm of each column of $\mathbf{W}^i$ (τ and w are well defined due to the properties of structured spinners). The last inequality comes from the fact that $\mathbf{x}$ and $\mathbf{y}$ are orthogonal. Now if we define matrices $\mathbf{A}^{i,j}$ as in the definition of the model of structured spinners then we see that

$$(\mathbf{v}^1)^T \cdot \mathbf{v}^2 = \sum_{l,u\in\{1,...,n\}} \rho_l\rho_u T^{i,j}_{l,u}, \quad (18)$$

where: $T^{i,j}_{l,u} = x_l y_u A^{i,j}_{l,u}$.

Now we will use the following inequality:

Theorem 6 (Hanson-Wright Inequality) *Let $\boldsymbol{X} = (X_1, ..., X_n)^T \in \mathbb{R}^n$ be a random vector with independent components X_i which satisfy: $\mathbb{E}[X_i] = 0$ and have sub-Gaussian norm at most K for some given $K > 0$. Let $\boldsymbol{A}$ be an $n \times n$ matrix. Then for every $t \geq 0$ the following is true:*

$$\mathbb{P}[\boldsymbol{X}^T\boldsymbol{A}\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}^T\boldsymbol{A}\boldsymbol{X}] > t] \leq 2e^{-c\min(\frac{t^2}{K^4\|\boldsymbol{A}\|_F^2}, \frac{t}{K^2\|\boldsymbol{A}\|_2})}, \quad (19)$$

where c is some universal positive constant.

Note that, assuming $\delta(n)$-balancedness, we have: $\|\mathbf{T}^{i,j}\|_F \leq \frac{\delta^2(n)}{n}\|\mathbf{A}^{i,j}\|_F$ and $\|\mathbf{T}^{i,j}\|_2 \leq \frac{\delta^2(n)}{n}\|\mathbf{A}^{i,j}\|_2$.

Now we take $\mathbf{X} = (\rho_1, ..., \rho_n)^T$ and $\mathbf{A} = \mathbf{T}^{i,j}$ in the theorem above. Applying the Hanson-Wright inequality in that setting, taking the union bound over all pairs of different vectors $\mathbf{v}^i, \mathbf{v}^j \in \mathcal{A}$ (this number is exactly: $\binom{md}{2}$) and the event $\mathcal{E}_{balanced}$, finally taking the union bound over all s functions f_i, we conclude that with probability at least:

$$p_{good} = 1 - p(n)ds - 2\binom{md}{2}se^{-\Omega(\min(\frac{t^2n^2}{K^4\Lambda_F^2\delta^4(n)}, \frac{tn}{K^2\Lambda_2\delta^2(n)}))} \quad (20)$$

for every f any two different vectors $\mathbf{v}^i, \mathbf{v}^j \in \mathcal{A}$ satisfy: $|(\mathbf{v}^i)^T \cdot \mathbf{v}^j| \leq t$.

Note that from the fact that $\mathbf{B}_2\mathbf{B}_1$ is $(\delta(n), p(n))$-balanced and from Equation 20, we get that with probability at least:

$$p_{right} = 1 - 2p(n)ds - 2\binom{md}{2}se^{-\Omega(\min(\frac{t^2n^2}{K^4\Lambda_F^2\delta^4(n)}, \frac{tn}{K^2\Lambda_2\delta^2(n)}))}. \quad (21)$$

for every f any two different vectors $\mathbf{v}^i, \mathbf{v}^j \in \mathcal{A}$ satisfy: $|(\mathbf{v}^i)^T \cdot \mathbf{v}^j| \leq t$ and furthermore each $\mathbf{v}^i$ is $\delta(n)$-balanced.

Assume now that this event happens. Consider the vector

$$\mathbf{q}' = ((\mathbf{y}^1)^T, ..., (\mathbf{y}^d)^T)^T \in \mathbb{R}^{md}. \quad (22)$$

Note that $\mathbf{q}'$ can be equivalently represented as:

$$\mathbf{q}' = (\mathbf{r}^T \cdot \mathbf{v}^1, ..., \mathbf{r}^T \cdot \mathbf{v}^{md}), \quad (23)$$

where: $\mathcal{A} = \{\mathbf{v}^1, ..., \mathbf{v}^{md}\}$. From the fact that $\phi_i\mathbf{B}_2$ and $\mathbf{B}_1$ are isometries we conclude that: $\|\mathbf{v}^i\|_2 = 1$ for $i = 1, ...$

Now we will need the following Berry-Esseen type result for random vectors:

Theorem 7 (Bentkus [Bentkus, 2003]) *Let $\boldsymbol{X}_1, ..., \boldsymbol{X}_n$ be independent vectors taken from $\mathbb{R}^k$ with common mean $\mathbb{E}[\boldsymbol{X}_i] = 0$. Let $\boldsymbol{S} = \boldsymbol{X}_1 + ... + \boldsymbol{X}_n$. Assume that the covariance operator $\boldsymbol{C}^2 = cov(\boldsymbol{S})$ is invertible. Denote $\beta_i = \mathbb{E}[\|\boldsymbol{C}^{-1}\boldsymbol{X}_i\|_2^3]$ and $\beta = \beta_1 + ... + \beta_n$. Let $\mathcal{C}$ be the set of all convex subsets of $\mathbb{R}^k$. Denote $\Delta(\mathcal{C}) = \sup_{A \in \mathcal{C}} |\mathbb{P}[S \in A] - \mathbb{P}[Z \in A]|$, where Z is the multivariate Gaussian distribution with mean 0 and covariance operator $\boldsymbol{C}^2$. Then:*

$$\Delta(\mathcal{C}) \leq ck^{\frac{1}{4}}\beta \tag{24}$$

for some universal constant c.

Denote: $\mathbf{X}_i = (r_i v_i^1, ..., r_i v_i^k)^T$ for $k = md$, $\mathbf{r} = (r_1, ..., r_n)^T$ and $\mathbf{v}^j = (v_1^j, ..., v_n^j)$. Note that $\mathbf{q}' = \mathbf{X}_1 + ... + \mathbf{X}_n$. Clearly we have: $\mathbb{E}[\mathbf{X}_i] = 0$ (the expectation is taken with respect to the random choice of $\mathbf{r}$). Furthermore, given the choices of $\mathbf{v}^1, ..., \mathbf{v}^k$, random vectors $\mathbf{X}_1, .., \mathbf{X}_n$ are independent.

Let us calculate now the covariance matrix of $\mathbf{q}'$. We have:

$$\mathbf{q}'_i = r_1 v_1^i + ... + r_n v_n^i, \tag{25}$$

where: $\mathbf{q}' = (\mathbf{q}'_1, ..., \mathbf{q}'_k)$.

Thus for i_1, i_2 we have:

$$\begin{aligned}\mathbb{E}[\mathbf{q}'_{i_1}\mathbf{q}'_{i_2}] &= \sum_{j=1}^{n} v_j^{i_1} v_j^{i_2} \mathbb{E}[r_j^2] + 2 \sum_{1 \leq j_1 < j_2 \leq n} v_{j_1}^{i_1} v_{j_2}^{i_2} \mathbb{E}[r_{j_1} r_{j_2}] \\ &= (\mathbf{v}^{i_1})^T \cdot \mathbf{v}^{i_2}, \end{aligned} \tag{26}$$

where the last equation comes from the fact r_j are either Gaussian from $\mathcal{N}(0,1)$ or discrete with entries from $\{-1, +1\}$ and furthermore different r_js are independent.

Therefore if $i_1 = i_2 = i$, since each $\mathbf{v}^i$ has unit L_2-norm, we have that

$$\mathbb{E}[\mathbf{q}'_i \mathbf{q}'_i] = 1, \tag{27}$$

and for $i_1 \neq i_2$ we get:

$$|\mathbb{E}[\mathbf{q}'_{i_1} \mathbf{q}'_{i_2}]| \leq t. \tag{28}$$

We conclude that the covariance matrix $\Sigma_{\mathbf{q}'}$ of the distribution $\mathbf{q}'$ is a matrix with entries 1 on the diagonal and other entries of absolute value at most t.

For $t = o_k(1)$ small enough and from the $\delta(n)$-balancedness of vectors $\mathbf{v}^1, ..., \mathbf{v}^k$ we can conclude that:

$$\mathbb{E}[\|\mathbf{C}^{-1}\mathbf{X}_i\|_2^3] = O(\mathbb{E}[\|\mathbf{X}_i\|_2^3]) = O(\sqrt{(\frac{k}{n})^3}\delta^3(n)), \tag{29}$$

Now, using Theorem 7, we conclude that

$$\begin{aligned}\sup_{A \in \mathcal{C}} |\mathbb{P}[\mathbf{q}' \in A] - \mathbb{P}[Z \in A]| &= O(k^{\frac{1}{4}} n \cdot \frac{k^{\frac{3}{2}}}{n^{\frac{3}{2}}}\delta^3(n)) \\ &= O(\frac{\delta^3(n)}{\sqrt{n}} k^{\frac{7}{4}}), \end{aligned} \tag{30}$$

where Z is taken from the multivariate Gaussian distribution with covariance matrix $\mathbf{I} + \mathbf{E}$ and $\mathcal{C}$ is the set of all convex sets. Now if we apply the above inequality to the pairwise disjoint convex sets $A_1, ..., A_j$, where $A_1 \cup ... \cup A_j = f_i^{-1}(\mathcal{S})$ and $l \leq b$ (such sets exist form the b-convexity of $f_i^{-1}(\mathcal{S})$), take $\eta = \frac{\delta^3(n)}{\sqrt{n}} k^{\frac{7}{4}}$, $\epsilon = t = o_{md}(1)$ and take n large enough, the statement of the theorem follows. □

5.4.5 Proof of Theorem 2

Proof: Let us assume that f_i is a convex function of $\mathbf{q}_{f_i}$ (if f_i is concave then the proof completely analogous). For any $t \in \mathbb{R}$ let $\mathcal{S}_t = \{\mathbf{q}_{f_i} : f_i(\mathbf{q}_{f_i}) \leq t\}$ for f_i and $\mathcal{S}_t = \{\mathbf{q}_{f'_i} : f'_i(\mathbf{q}_{f'_i}) \leq t\}$ for f'_i. From the convexity assumption we get that $\mathcal{S}_t$ is a convex set. Thus we can directly apply Theorem 1 and the result regarding cdf functions follows. To obtain the result regarding the characteristic functions, notice first that we have:

$$\phi_X(t) = \int_{-1}^{1} \mathbb{P}[cos(tX) > s]ds + i \int_{-1}^{1} \mathbb{P}[sin(tX) > s]ds \tag{31}$$

The event $\{cos(tX) > s\}$ for $t \neq 0$ is equivalent to: $X \in \cup_{I \in \mathcal{I}} I$ for some family of intervals $\mathcal{I}$. Similar observation is true for the event $\{sin(tX) > s\}$.

In our scenario, from the fact that f_i is bounded, we conclude that the corresponding families $\mathcal{I}$ are finite. Furthermore, the probability of belonging to a particular interval can be expressed by the values of the cdf function in the endpoints of that interval. From this observation and the result on cdfs that we have just obtained, the result for the characteristic functions follows immediately. □

5.4.6 Proof of Theorem 3

Proof: This comes directly from Theorem 1 and Lemma 1. □

5.4.7 Proof of Theorem 4

Proof: For clarity we will assume that the structured matrix consists of just one block of m rows and will compare its performance with the unstructured variant of m rows (the more general case when the structured matrix is obtained by stacking vertically many blocks is analogous since the blocks are chosen independently).

Consider the two-dimensional linear space $\mathcal{H}$ spanned by $\mathbf{x}$ and $\mathbf{y}$. Fix some orthonormal basis $\mathcal{B} = \{\mathbf{u}^1, \mathbf{u}^2\}$ of $\mathcal{H}$. Take vectors $\mathbf{q}$ and $\mathbf{q}'$. Note that they are $2m$-dimensional, where m is the number of rows of the block used in the structured setting. From Theorem 3 we conclude that will probability at least $p_{success}$, where $p_{success}$ is as in the statement of the theorem the following holds for any convex $2m$-dimensional set

A:

$$|\mathbb{P}[\mathbf{q}(\epsilon) \in A] - \mathbb{P}[\mathbf{q}' \in A]| \leq \eta, \quad (32)$$

where $\eta = \frac{\log^3(n)}{n^{\frac{2}{5}}}$. Take two corresponding entries of vectors $\mathbf{v}^1_{\mathbf{x},\mathbf{y}}$ and $\mathbf{v}^2_{\mathbf{x},\mathbf{y}}$ indexed by a pair $(\mathbf{e}_i, \mathbf{e}_j)$ for some fixed $i, j \in \{1, ..., m\}$ (for the case when the pair is not of the form $(\mathbf{e}, \mathbf{e}_j)$, but of a general form: $(\pm\mathbf{e}_i, \pm\mathbf{e}_j)$ the analysis is exactly the same). Call them p^1 and p^2 respectively. Our goal is to compute $|p^1 - p^2|$. Notice that p^1 is the probability that $h(\mathbf{x}) = \mathbf{e}_i$ and $h(\mathbf{y}) = \mathbf{e}_j$ for the unstructured setting and p^2 is that probability for the structured variant.

Let us consider now the event $E^1 = \{h(\mathbf{x}) = \mathbf{e}_i \wedge h(\mathbf{y}) = \mathbf{e}_j\}$, where the setting is unstructured. Denote the corresponding event for the structured setting as E^2. Denote $\mathbf{q} = (q_1, ..., q_{2m})$. Assume that $\mathbf{x} = \alpha_1\mathbf{u}^1 + \alpha_2\mathbf{u}^2$ for some scalars $\alpha_1, \alpha_2 > 0$. Denote the unstructured Gaussian matrix by $\mathbf{G}$. We have:

$$\mathbf{G}\mathbf{x} = \alpha_1\mathbf{G}\mathbf{u}^1 + \alpha_2\mathbf{G}\mathbf{u}^2 \quad (33)$$

Note that we have: $\mathbf{G}\mathbf{u}^1 = (q_1, ..., q_m)^T$ and $\mathbf{G}\mathbf{u}^2 = (q_{m+1}, ..., q_{2m})^T$. Denote by $A(\mathbf{e}_i)$ the set of all the points in $\mathbb{R}^m$ such that their angular distance to $\mathbf{e}_i$ is at most the angular distance to all other $m-1$ canonical vectors. Note that this is definitely the convex set. Now denote:

$$Q(\mathbf{e}_i) = \{(q_1, ..., q_{2m})^T \in \mathbb{R}^{2m} : \alpha_1(q_1, ..., q_m)^T + \alpha_2(q_{m+1}, ..., q_{2m})^T \in A(\mathbf{e}_i)\}. \quad (34)$$

Note that since $A(\mathbf{e}_i)$ is convex, we can conclude that $Q(\mathbf{e}_i)$ is also convex. Note that

$$\{h(\mathbf{x}) = \mathbf{e}_i\} = \{\mathbf{q} \in Q(\mathbf{e}_i)\}. \quad (35)$$

By repeating the analysis for the event $\{h(\mathbf{y}) = \mathbf{e}_j\}$, we conclude that:

$$\{h(\mathbf{x}) = \mathbf{e}_i \wedge h(\mathbf{y}) = \mathbf{e}_j\} = \{\mathbf{q} \in Y(\mathbf{e}_i, \mathbf{e}_j)\} \quad (36)$$

for convex set $Y(\mathbf{e}_i, \mathbf{e}_j) = Q(\mathbf{e}_i) \cap Q(\mathbf{e}_j)$. Now observe that

$$|p^1 - p^2| = |\mathbb{P}[\mathbf{q} \in Y(\mathbf{e}_i, \mathbf{e}_j)] - \mathbb{P}[\mathbf{q}' \in Y(\mathbf{e}_i, \mathbf{e}_j)]| \quad (37)$$

Thus we have:

$$|p^1 - p^2| \leq |\mathbb{P}[\mathbf{q} \in Y(\mathbf{e}_i, \mathbf{e}_j)] - \mathbb{P}[\mathbf{q}(\epsilon) \in Y(\mathbf{e}_i, \mathbf{e}_j)]| + |\mathbb{P}[\mathbf{q}(\epsilon) \in Y(\mathbf{e}_i, \mathbf{e}_j)] - \mathbb{P}[\mathbf{q}' \in Y(\mathbf{e}_i, \mathbf{e}_j)]| \quad (38)$$

Therefore we have:

$$|p^1 - p^2| \leq |\mathbb{P}[\mathbf{q} \in Y(\mathbf{e}_i, \mathbf{e}_j)] - \mathbb{P}[\mathbf{q}(\epsilon) \in Y(\mathbf{e}_i, \mathbf{e}_j)]| + \eta. \quad (39)$$

Thus we just need to upper-bound:

$$\xi = |\mathbb{P}[\mathbf{q} \in Y(\mathbf{e}_i, \mathbf{e}_j)] - \mathbb{P}[\mathbf{q}(\epsilon) \in Y(\mathbf{e}_i, \mathbf{e}_j)]|. \quad (40)$$

Denote the covariance matrix of the distribution $\mathbf{q}(\epsilon)$ as $\mathbf{I} + \mathbf{E}$. Note that $\mathbf{E}$ is equal to 0 on the diagonal and the absolute value of all other off-diagonal entries is at most ϵ.

Denote $k = 2m$. We have

$$\xi = |A - B|,$$

where $A = \frac{1}{(2\pi)^{\frac{k}{2}}\sqrt{\det(I+E)}} \int_{Y(\mathbf{e}_i,\mathbf{e}_j)} e^{-\frac{\mathbf{x}^T(\mathbf{I}+\mathbf{E})^{-1}\mathbf{x}}{2}} d\mathbf{x}$

and $B = \frac{1}{(2\pi)^{\frac{k}{2}}} \int_{Y(\mathbf{e}_i,\mathbf{e}_j)} e^{-\frac{\mathbf{x}^T\mathbf{x}}{2}} d\mathbf{x}$.

Expanding: $(\mathbf{I} + \mathbf{E})^{-1} = \mathbf{I} - \mathbf{E} + \mathbf{E}^2 - ...$, noticing that $|\det(I + E) - 1| = O(\epsilon^{2m})$, and using the above formula, we easily get:

$$\xi = O(\epsilon). \quad (41)$$

That completes the proof. □

5.4.8 b-convexity for angular kernel approximation

Let us now consider the setting, where linear projections are used to approximate angular kernels between paris of vectors via random feature maps. In this case, the linear projection is followed by the pointwise nonlinear mapping, where the applied nonlinear mapping is a sign function. The angular kernel is retrieved from the Hamming distance between $\{-1, +1\}$-hashes obtained in such a way. Note that in this case we can assign to each pair $\mathbf{x}, \mathbf{y}$ of vectors from a database a function $f_{\mathbf{x},\mathbf{y}}$ that outputs the binary vector which length is the size of the hash and with these indices turned on for which the hashes of $\mathbf{x}$ and $\mathbf{y}$ disagree. Such a binary vector uniquely determines the Hadamard distance between the hashes. Notice that for a fixed-length hash $f_{\mathbf{x},\mathbf{y}}$ produces only finitely many outputs. If $\mathcal{S}$ is a set-singleton consisting of one of the possible outputs, then one can notice (straightforwardly from the way the hash is created) that $f^{-1}_{\mathbf{x},\mathbf{y}}(\mathcal{S})$ is an intersection of the convex sets (as a function of $\mathbf{q}_{f_{\mathbf{x},\mathbf{y}}}$). Thus it is convex and thus for sets $\mathcal{S}$ which are singletons we can take $b = 1$.

5.4.9 Proof of Theorem 5

In this section, we show that by learning vector $\mathbf{r} \in \mathbb{R}^k$ from the definition above, one can approximate well any matrix $\mathbf{M} \in \mathbb{R}^{m\times n}$ learned by the neural network, providing that the size k or $\mathbf{r}$ is large enough in comparison with the number of projections and the intrinsic dimensionality d of the data $\mathcal{X}$.

Take the parametrized structured spinner matrix $\mathbf{M}_{struct} \in \mathbb{R}^{m\times n}$ with a learnable vector $\mathbf{r}$. Let $\mathbf{M} \in \mathbb{R}^{m\times n}$ be a matrix learned in the unstructured setting.

Let $\mathcal{B} = \{\mathbf{x}^1, ..., \mathbf{x}^d\}$ be some orthonormal basis of the linear space, where data $\mathcal{X}$ is taken from.

Proof: Note that from the definition of the parametrized structured spinner model we can conclude that with probability at least $p_1 = 1 - p(n)$ with respect to the choices of $\mathbf{M}_1$ and $\mathbf{M}_2$ each $\mathbf{M}_{struct}\mathbf{x}^i$ is of the form:

$$\mathbf{M}_{struct}\mathbf{x}^i = (\mathbf{r}^T \cdot \mathbf{z}_1(\mathbf{q}^i), ..., \mathbf{r}^T \cdot \mathbf{z}_m(\mathbf{q}^i))^T, \quad (42)$$

where each $\mathbf{z}_j(\mathbf{q}^i)$ is of the form:

$$\mathbf{z}_j(\mathbf{q}^i) = (w^j_{1,1}\rho_1 q^i_1 + w^j_{1,n}\rho_n q^i_n, ..., w^j_{k,1}\rho_1 q^i_1 + w^j_{k,n}\rho_n q^i_n)^T \quad (43)$$

and $\mathcal{B}' = \{\mathbf{q}^1, ..., \mathbf{q}^d\}$ is an orthonormal basis such that: $\|\mathbf{q}^i\|_\infty \leq \frac{\delta(n)}{\sqrt{n}}$ for $i = 1, ..., n$.

Note that the system of equations:

$$\mathbf{M}^{struct}\mathbf{x}^i = \mathbf{M}\mathbf{x}^i \quad (44)$$

for $i = 1, ..., d$ has the solution in $\mathbf{r}$ if the vectors from the set $\mathcal{A} = \{\mathbf{z}_j(\mathbf{q}^i) : j = 1, ..., m, i = 1, ...d\}$ are independent.

Construct a matrix $\mathbf{G} \in \mathbb{R}^{md \times k}$, where rows are vectors from $\mathcal{A}$. We want to show that $rank(\mathbf{G}) = md$. It suffices to show that $det(\mathbf{G}\mathbf{G}^T) \neq 0$. Denote $\mathbf{B} = \mathbf{G}\mathbf{G}^T$. Note that $B_{i,j} = (\mathbf{v}^i)^T\mathbf{v}^j$, where $\mathcal{A} = \{\mathbf{v}^1, ..., \mathbf{v}^{md}\}$. Take two vectors $\mathbf{v}^a, \mathbf{v}^b \in \mathcal{A}$. Note that from the definition of $\mathcal{A}$ we get:

$$(\mathbf{v}^a)^T\mathbf{v}^b = \sum_{l \in \{1,...,n\}, u \in \{1,...,n\}} \rho_l \rho_u x_l y_u (\sum_{s=1}^{k} w^i_{s,l} w^j_{s,u}) \quad (45)$$

for some i, j and some vectors $\mathbf{x} = (x_1, ..., x_n)^T$, $\mathbf{y} = (y_1, ..., y_n)^T$. Furthermore,

- $i = j$ and $\mathbf{x} = \mathbf{y}$ if $a = b$,
- $\|\mathbf{x}\|_2 = \|\mathbf{y}\|_2 = 1$,
- $\mathbf{x}^T\mathbf{y} = 0$ or $\mathbf{x} = \mathbf{y}$ and $i \neq j$ for $a \neq b$.

We also have:

$$\mathbb{E}[(\mathbf{v}^a)^T\mathbf{v}^b] = \mathbb{E}[\sum_{l \in \{1,...,n\}} \rho_l^2 x_l y_l (\sum_{s=1}^{k} w^i_{s,l} w^j_{s,u})]. \quad (46)$$

From the previous observations and the properties of matrices $\mathbf{W}^1, ..., \mathbf{W}^n$ we conclude that the entries of the diagonal of $\mathbf{B}$ are equal to 1. Furthermore, all other entries are 0 on expectation. Using Hanson-Wright inequality, we conclude that for any $t > 0$ we have: $|B_{i,j}| \leq t$ for all $i \neq j$ with probability at least:

$$p_{succ} = 1 - 2p(n)d - 2\binom{md}{2} e^{-c\min(\frac{t^2 n^2}{K^4 \Lambda_F^2 \delta^4(n)}, \frac{tn}{K^2 \Lambda_2 \delta^2(n)})}.$$

If this is the case, we let $\tilde{\mathbf{B}} \in \mathbb{R}^{(md)\times(md)}$ be a matrix with diagonal entries $\tilde{\mathbf{B}}_{i,i} = 0$ and off-diagonal entries $\tilde{\mathbf{B}}_{\mathbf{i,j}} = -\mathbf{B}_{i,j}$. Furthermore, let $\mathbf{B}^* \in \mathbb{R}^{(md)\times(md)}$ be a matrix with diagonal entries $\mathbf{B}^*_{i,i} = 0$ and off-diagonal entries $\mathbf{B}^*_{i,j} = t$.

Following a similar argument as in [Brent et al., 2014], note that $\mathbf{B}^* = t(\mathbf{J} - \mathbf{I})$ where $\mathbf{J}$ is the matrix of all ones (thus of rank 1) and $\mathbf{I}$ is the identity matrix. Then the eigenvalues of $\mathbf{B}^*$ are $t(md-1)$ with multiplicity 1 and $t(0-1)$ with multiplicity $(md-1)$. We, thereby, are able to explicitly compute $\det(\mathbf{I} - \mathbf{B}^*) = (1 - t(md-1))(1+t)^{md-1}$.

If $\rho(\mathbf{B}^*) \leq 1$, we can apply Theorem 1 of [Brent et al., 2014] by replacing $\mathbf{F}$ with $\mathbf{B}^*$ and $\mathbf{E}$ with $\tilde{\mathbf{B}}$. For the convenience of the reader, we state their theorem here: Let $\mathbf{F} \in \mathbb{R}^{n\times n}$ with non-negative entries and $\rho(F) \leq 1$. Let $\mathbf{E} \in \mathbb{R}^{n\times n}$ with entries $\mid e_{i,j} \mid \leq f_{i,j}$, then $\det(\mathbf{I} - \mathbf{E}) \geq \det(\mathbf{I} - \mathbf{F})$.

That is: if $\rho(\mathbf{B}^*) \leq 1$, then

$$det(\mathbf{I} - \mathbf{B}^*) = (1 - t(md-1))(1+t)^{md-1} \leq det(\mathbf{I} - \tilde{\mathbf{B}}) = det(\mathbf{B}). \quad (47)$$

The final step is to observe that:
$\rho(\mathbf{B}^*) \leq 1 \iff \max\{\mid t(md-1) \mid, \mid -t \mid\} = t(md-1) \leq 1 \iff t \leq \frac{1}{md-1}$. Using this result, we, hence, see that $det(\mathbf{B}) \geq (1 - t(md-1))(1+t)^{md-1} \geq 0$, in particular $det(\mathbf{B}) > 0$ for $t = \frac{1}{md}$. That completes the proof. □

5.4.10 Additional experiments

This experiment focuses on the Newton sketch approach [Pilanci and Wainwright, 2015], a generic optimization framework. It guarantees super-linear convergence with exponentially high probability for self-concordant functions, and a reduced computational complexity compared to the original second-order Newton method. The method relies on using a sketched version of the Hessian matrix, in place of the original one. In the subsequent experiment we show that matrices from the family of strucured spinners can be used for this purpose, thus can speed up several convex optimization problems solvers.

We consider the unconstrained large scale logistic regression problem, i.e. given a set of n observations $\{(a_i, y_i)\}_{i=1..n}$, with $a_i \in \mathbb{R}^d$ and $y_i \in \{-1, 1\}$, find

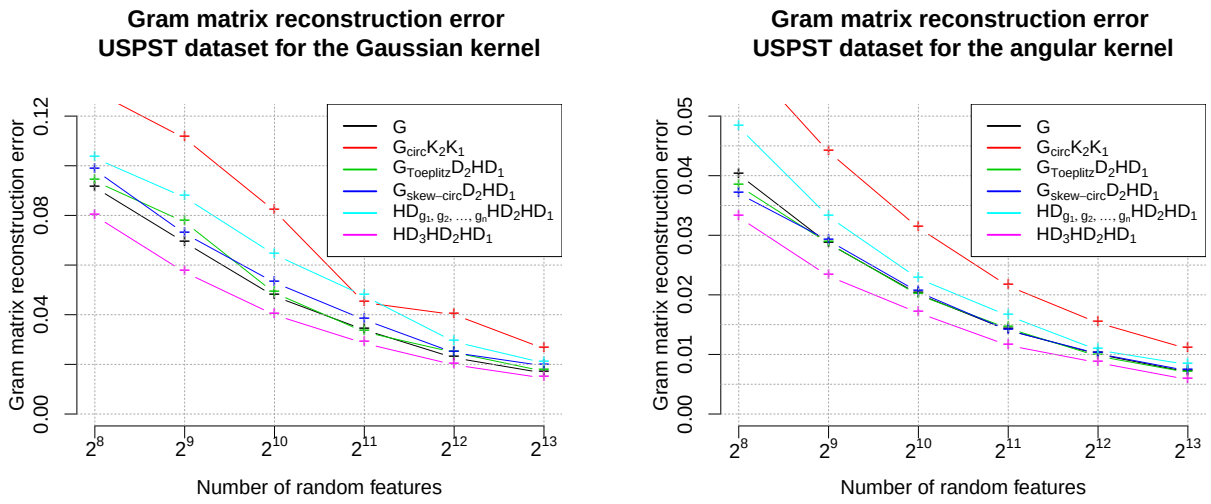

Figure 5: Accuracy of random feature map kernel approximation for the USPST dataset.

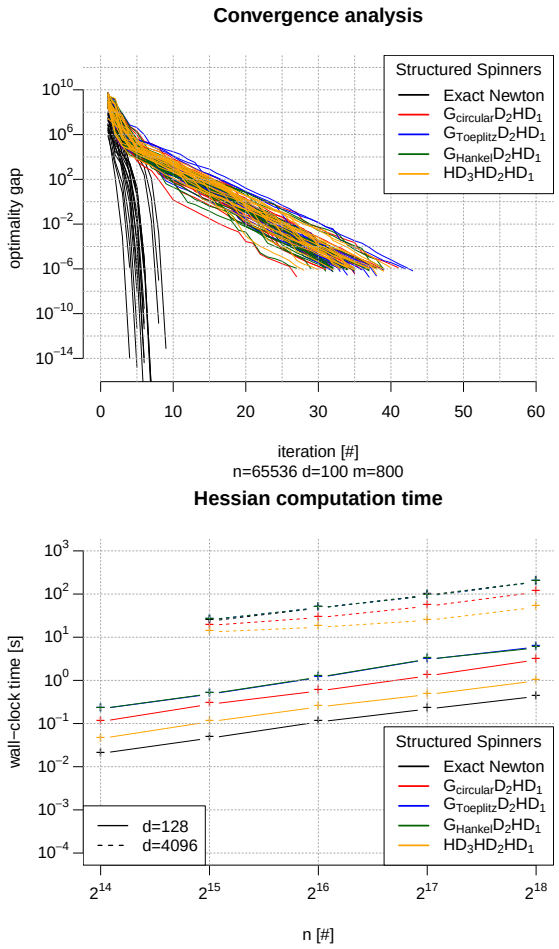

Figure 6: Numerical illustration of the convergence (top) and computational complexity (bottom) of the Newton sketch algorithm with various structured spinners. (left) Various sketching structures are compared in terms of the convergence against iteration number. (bottom) Wall-clock times of structured spinners are compared in various dimensionality settings.

$x \in \mathbb{R}^d$ minimizing the cost function

$$f(x) = \sum_{i=1}^{n} \log(1 + \exp(-y_i a_i^T x)) \ .$$

The Newton approach to solving this optimization problem entails solving at each iteration the least squares equation $\nabla^2 f(x^t)\Delta^t = -\nabla f(x^t)$, where

$$\nabla^2 f(x^t) = A^T \mathrm{diag}\left(\frac{1}{1+\exp(-a_i^T x)}(1 - \frac{1}{1+\exp(-a_i^T x)})\right) A \in \mathbb{R}^{d\times d}$$

is the Hessian matrix of $f(x^t)$, $A = [a_1^T a_2^T \cdots a_n^T] \in \mathbb{R}^{n\times d}$, $\Delta^t = x^{t+1} - x^t$ is the increment at iteration t and $\nabla f(x^t) \in \mathbb{R}^d$ is the gradient of the cost function. In [Pilanci and Wainwright, 2015] it is proposed to consider the sketched version of the least square equation, based on a Hessian square root of $\nabla^2 f(x^t)$, denoted $\nabla^2 f(x^t)^{1/2} = \mathrm{diag}\left(\frac{1}{1+\exp(-a_i^T x)}(1 - \frac{1}{1+\exp(-a_i^T x)})\right)^{1/2} A \in \mathbb{R}^{n\times d}$. The least squares problem at each iteration t is of the form:

$$\left((S^t \nabla^2 f(x^t)^{1/2})^T S^t \nabla^2 f(x^t)^{1/2}\right) \Delta^t = -\nabla f(x^t) \ ,$$

where $S^t \in \mathbb{R}^{m\times n}$ is a sequence of isotropic sketch matrices. Let's finally recall that the gradient of the cost function is

$$\nabla f(x^t) = \sum_{i=1}^{n} \left(\frac{1}{1+\exp(-y_i a_i^T x)} - 1\right) y_i a_i \ .$$

In our experiment, the goal is to find $x \in \mathbb{R}^d$, which minimizes the logistic regression cost, given a dataset

$\{(a_i, y_i)\}_{i=1..n}$, with $a_i \in \mathbb{R}^d$ sampled according to a Gaussian centered multivariate distribution with covariance $\Sigma_{i,j} = 0.99^{|i-j|}$ and $y_i \in \{-1, 1\}$, generated at random. Various sketching matrices $S^t \in \mathbb{R}^{m \times n}$ are considered.

In Figure 6 we report the convergence of the Newton sketch algorithm, as measured by the optimality gap defined in [Pilanci and Wainwright, 2015], versus the iteration number. As expected, the structured sketched versions of the algorithm do not converge as quickly as the exact Newton-sketch approach, however various matrices from the family of structured spinners exhibit equivalent convergence properties as shown in the figure.

When the dimensionality of the problem increases, the cost of computing the Hessian in the exact Newton-sketch approach becomes very large [Pilanci and Wainwright, 2015], scaling as $\mathcal{O}(nd^2)$. The complexity of the structured Newton-sketch approach with the matrices from the family of structured spinners is instead only $\mathcal{O}(dn\log(n) + md^2)$. Figure 6 also illustrates the wall-clock times of computing single Hessian matrices and confirms that the increase in number of iterations of the Newton sketch compared to the exact sketch is compensated by the efficiency of sketched computations, in particular Hadamard-based sketches yield improvements at the lowest dimensions.