Supplementary material

A.1 Properties of LogDet subproblem

Proof of Theorem 3.2

The optimality condition of (11) is $P_{\mathcal{G}}(X^{-1}) = P_{\mathcal{G}}(H)$, $X \in S_n^{++}(\mathcal{G})$. Let $Z = L^{-T}D^{-1}L^{-1}$, then $P_{\mathcal{G}}(Z) = H$

$$ZL = L^{-T}D^{-1} \implies ZLe_j = L^{-T}D^{-1}e_j$$

Let $J_j = I_j \cup j$, where $I_j = \{j+1, \ldots, j+b\}$ as defined in the theorem, then select J_j indices of vectors on both sides of the second equality above and selecting the J_j indices :

$$\begin{bmatrix} Z_{jj} & Z_{jI_j} \\ Z_{I_jj} & Z_{J_jJ_j} \end{bmatrix} \begin{bmatrix} 1 \\ L_{I_j} \end{bmatrix} = \begin{bmatrix} 1/d_{jj} \\ 0 \end{bmatrix} \tag{15}$$

Note that L^{-T} is an upper triangular matrix with ones in the diagonal hence J_j^{th} block of $L^{-T}e_j$ will be $[1, 0, 0, \ldots]$. Also, since $P_{\mathcal{G}}(Z) = H$

$$\begin{bmatrix} Z_{jj} & Z_{jI_j} \\ Z_{I_jj} & Z_{J_jJ_j} \end{bmatrix} = \begin{bmatrix} H_{jj} & H_{jI_j} \\ H_{I_jj} & H_{J_jJ_j} \end{bmatrix}$$

Substituting this in the linear equation 15

$$\begin{bmatrix} H_{jj} & H_{jI_j} \\ H_{I_jj} & H_{J_jJ_j} \end{bmatrix} \begin{bmatrix} 1 \\ L_{I_j} \end{bmatrix} = \begin{bmatrix} 1/d_{jj} \\ 0 \end{bmatrix}$$

$$\begin{bmatrix} H_{jj} & H_{jI_j} \\ H_{I_jj} & H_{J_jJ_j} \end{bmatrix} \begin{bmatrix} d_{jj} \\ d_{jj} \cdot L_{I_j} \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

$$H_{jj}d_{jj} + d_{jj}H_{I_jj}^T L_{I_jj} = 1$$

$$H_{I_jj}d_{jj} + d_{jj}H_{I_jI_j}L_{I_jj} = 0$$

The lemma follows from solving the above equations. Note that here we used that lower triangular halves of matrices L and H have the same sparsity patterns, which follows from the fact that banded graph is a chordal graph with perfect elimination order $\{1, 2, \ldots, n\}$. Furthermore, X_t is positive definite, since as $(H_{jj} - H_{I_j j}^T H_{I_j I_j}^{-1} H_{I_j j})$ is a schur complement of submatrix of H formed by $J_j = I_j \cup \{j\}$.

Proof of Theorem 3.1 The proof follows trivially from Theorem 3.1, when b is set to 1.

A.2 Regret bound analysis

Proof sketch of Theorem 3.3. We decompose the regret into $R_T \leq T_1 + T_2 + T_3$ in Lemma .1 and individually bound the terms. Term $T_2 = \frac{1}{2\eta} \cdot \sum_{t=1}^{T-1} (w_{t+1} - w^*)^T (X_{t+1}^{-1} - X_t^{-1})(w_{t+1} - w^*)$ depends on closeness of consecutive inverses of preconditioners, $(X_{t+1}^{-1} - X_t^{-1})$, to upperbound this we first give explicit expressions of X_t^{-1} for tridiagonal preconditioner in Lemma .2 in Appendix A.2.2. This explicit expression is later used to bound each entry of $(X_{t+1}^{-1} - X_t^{-1})$ with $O(1/\sqrt{t})$ in Appendix A.2.4, this gives a $O(\sqrt{T})$ upperbound on T_2. To show an upperbound on $T_3 = \sum_{t=1}^{T} \frac{\eta}{2} \cdot g_t^T X_t g_t$, we individually bound $g_t^T X_t g_t$ by using a Loewner order $X_t \preceq \|X_t\|_2 I_n \preceq \|X_t\|_\infty I_n$ and show that $\|X_t\|_\infty = \mathcal{O}(1/\sqrt{T})$ and consequently $T_3 = \mathcal{O}(\sqrt{T})$.

A.2.1 Regret bound decomposition

In this subsection we state Lemma .1 which upper bound the regret R_T using three terms T_1, T_2, T_3.

Lemma .1 ([25]). *In the OCO problem setup, if a prediction $w_t \in \mathbb{R}^n$ is made at round t and is updated as $w_{t+1} \coloneqq w_t - \eta X_t g_t$ using a preconditioner matrix $X_t \in S_n^{++}$*

$$R_T \leq \frac{1}{2\eta} \cdot (\|w_1 - w^*\|_{X_1^{-1}}^2 - \|w_{T+1} - w^*\|_{X_T^{-1}}) \tag{16}$$

$$+ \frac{1}{2\eta} \cdot \sum_{t=1}^{T-1} (w_{t+1} - w^*)^T (X_{t+1}^{-1} - X_t^{-1})(w_{t+1} - w^*) \tag{17}$$

$$+ \sum_{t=1}^{T} \frac{\eta}{2} \cdot g_t^T X_t g_t \tag{18}$$

Proof.

$$\begin{aligned}
\|w_{t+1} - w^*\|_{X_t^{-1}}^2 &= \|w_t - \eta X_t g_t - w^*\|_{X_t^{-1}}^2 \\
&= \|w_t - w^*\|_{X_t^{-1}}^2 + \eta^2 g_t^T X_t g_t \\
&\quad - 2\eta (w_t - w^*)^T g_t \\
\implies 2\eta (w_t - w^*)^T g_t &= \|w_t - w^*\|_{X_t^{-1}}^2 - \|w_{t+1} - w^*\|_{X_t^{-1}}^2 \\
&\quad + \eta^2 g_t^T X_t g_t
\end{aligned}$$

□

Using the convexity of f_t, $f_t(w_t) - f_t(w^*) \leq (w_t - w^*)^T g_t$, where $g_t = \Delta f_t(w_t)$ and summing over $t \in [T]$

$$R_T \leq \sum_{t=1}^{T} \frac{1}{2\eta} \cdot \left(\|w_t - w^*\|_{X_t^{-1}}^2 - \|w_{t+1} - w^*\|_{X_t^{-1}}^2 \right) \tag{19}$$

$$+ \frac{\eta}{2} \cdot g_t^T X_t g_t \tag{20}$$

The first summation can be decomposed as follows

$$\sum_{t=1}^{T}\left(\|w_t - w^*\|^2_{X_t^{-1}} - \|w_{t+1} - w^*\|^2_{X_t^{-1}}\right)$$
$$= \left(\|w_1 - w^*\|^2_{X_1^{-1}} - \|w_{T+1} - w^*\|^2_{X_T^{-1}}\right)$$
$$+ \sum_{t=1}^{T-1}(w_{t+1} - w^*)^T(X_{t+1}^{-1} - X_t^{-1})(w_{t+1} - w^*)$$

Substituting the above identity in the Equation (19) proves the lemma.

Let $R_T \leq T_1 + T_2 + T_3$, where

- $T_1 = \frac{1}{2\eta} \cdot (\|w_1 - w^*\|^2_{X_1^{-1}} - \|w_{T+1} - w^*\|_{X_T^{-1}})$
-
$$T_2 = \frac{1}{2\eta} \cdot \sum_{t=1}^{T-1}(w_{t+1} - w^*)^T(X_{t+1}^{-1} - X_t^{-1})(w_{t+1} - w^*) \tag{21}$$
- $T_3 = \sum_{t=1}^{T} \frac{\eta}{2} \cdot g_t^T X_t g_t$

A.2.2 Properties of tridiagonal preconditioner

In this subsection, we derive properties of the tridigonal preconditioner obtained from solving the LogDet subproblem (11) with $\mathcal{G}$ set to a chain graph over ordered set of vertices $\{1, \ldots, n\}$:

$$X_t = \underset{X \in S_n(\mathcal{G})^{++}}{\arg\min} \; -\log\det(X) + \mathrm{Tr}(XH_t) \tag{22}$$
$$= \underset{X \in S_n(\mathcal{G})^{++}}{\arg\min} \; \mathrm{D}_{\ell\mathrm{d}}(X, H_t^{-1}) \tag{23}$$

The second equality holds true only when H_t is positive definite. Although in Algorithm 1 we maintain a sparse $H_t = H_{t-1} + P_{\mathcal{G}}(g_t g_t^T/\lambda_t)$, $H_0 = \epsilon I_n$ which is further used in (22) to find the preconditioner X_t, our analysis assumes the full update $H_t = H_{t-1} + g_t g_t^T/\lambda_t$, $H_0 = \epsilon I_n$ followed by preconditioner X_t computation using (23). Note that the preconditioners X_t generated both ways are the same, as shown in Section 3.2.

The following lemma shows that the inverse of tridiagonal preconditioners used in Algorithm 1, will restore $H_{i,j}$, when (i, j) fall in the tridiagonal graph, else, the expression is related to product of $H_{i+k,i+k+1}$ corresponding to the edges in the path from node i to j in chain graph. This lemma will be used later in upperbounding T_2.

Lemma .2 (*Inverse of tridiagonal preconditioner*)**.** *If* $\mathcal{G}$ *= chain/tridiagonal graph and* $\hat{X} = \arg\min_{X \in S_n(\mathcal{G})^{++}} \mathrm{D}_{\ell\mathrm{d}}(X, H^{-1})$*, then the inverse* $\hat{X}^{-1}$ *has the following expression*

$$(\hat{X}^{-1})_{ij} = \begin{cases} H_{ij} & |i - j| \leq 1 \\ \frac{H_{ii+1}H_{i+1i+2}\ldots H_{j-1j}}{H_{i+1i+1}\ldots H_{j-1j-1}} & \end{cases} \tag{24}$$

Proof.

$$\hat{X}^{-1}\hat{X}^{(j)} = e_j$$

Where $\hat{X}^{(j)}$ is the j^{th} column of $\hat{X}$. Let $\hat{Y}$ denote the right hand side of Equation (24).

$$(\hat{Y}\hat{X})_{jj} = \hat{X}_{jj}\hat{Y}_{jj} + \hat{X}_{j-1j}\hat{Y}_{j-1j} + \hat{X}_{jj+1}\hat{Y}_{jj+1}$$
$$= \hat{X}_{jj}H_{jj} + \hat{X}_{j-1j}H_{j-1j} + \hat{X}_{jj+1}H_{jj+1}$$
$$= 1$$

The third equality is by using the following alternative form of Equation (12):

$$(\hat{X}^{(1)})_{i,j} = \begin{cases} 0, \text{ if } j - i > 1 \\ \frac{-H_{i,i+1}}{(H_{ii}H_{i+1,i+1} - H_{i+1,i+1}^2)}, \text{ if } j = i+1 \\ \frac{1}{H_{ii}}\left(1 + \sum_{j \in \text{neig}_{\mathcal{G}}(i)} \frac{H_{ij}^2}{H_{ii}H_{jj} - H_{ij}^2}\right), \text{ if } i = j \end{cases}, \tag{25}$$

where $i < j$. Similarly, the offdiagonals of $\hat{Y}\hat{X}$ can be evaluated to be zero as follows.

$$\begin{aligned} (\hat{Y}\hat{X})_{ij} &= \hat{Y}_{ij}\hat{X_j}j + \hat{Y}_{ij-1}\hat{X}_{j-1j} + \hat{Y}_{ij+1}\hat{X}_{j+1j} \\ &= \hat{Y}_{ij}\hat{X}_{jj} + \hat{Y}_{ij}\frac{H_{j-1j-1}}{H_{j-1j}} + \hat{Y}_{ij}\frac{H_{jj+1}}{H_{jj}}\hat{X}_{j+1j} \\ &= 0 \end{aligned}$$

□

Lemma .3. *Let $y \in \mathbb{R}^n$, $\beta = \max_t \max_{i \in [n-1]} |(H_t)_{ii+1}| / \sqrt{(H_t)_{ii}(H_t)_{i+1i+1}} < 1$, then*

$$y^T X_t^{-1} y \leq \|y\|_2^2 \|\text{diag}(H_t)\|_2 \left(\frac{1+\beta}{1-\beta}\right),$$

where X_t is defined as in Lemma .2.

Proof. Let $\tilde{X}_t^{-1} = \text{diag}(H_t)^{-1/2}\hat{X}_t \, \text{diag}(H_t)^{-1/2}$

$$y^T X_t^{-1} y \leq \left\|\text{diag}(H_t)^{1/2} y\right\|_2^2 \left\|\tilde{X}_t^{-1}\right\|_2 \tag{26}$$

Using the identity of spectral radius $\rho(X) \leq \|X\|_\infty$ and since $\tilde{X}$ is positive definite, $\left\|\tilde{X}_t^{-1}\right\|_2 \leq \|\tilde{X}_t^{-1}\|_\infty$

$$\begin{aligned} \left\|\tilde{X}_t^{-1}\right\|_2 &\leq \max_i \left\{\sum_j \left|(\tilde{X}_t^{-1})_{ij}\right|\right\} \\ &\leq 1 + 2(\beta + \beta^2 + \ldots) \\ &\leq \frac{1+\beta}{1-\beta} \end{aligned}$$

The second inequality is using Lemma .2. Substituting this in Equation (26) will give the lemma. □

A.2.3 Upperbounding Regret

The following Lemma is used in upperbounding both T_1 and T_3. In next subsection, we'll upper bound T_2 as well.

Lemma .4. *Let $\beta = \max_{t \in [T]} \max_{i \in [n-1]} |(H_t)_{ii+1}| / \sqrt{(H_t)_{ii}(H_t)_{i+1i+1}}$, then*

$$1/(1-\beta) \leq 8/\hat{\epsilon}^2,$$

where, $\hat{\epsilon}$ is a constant in parameter $\epsilon = \hat{\epsilon} G_\infty \sqrt{T}$ and consequently used in initializing $H_0 = \epsilon I_n$ in line 1 in Algorithm 1.

Proof.

$$1/(1-\beta) = \max_t \max_{i\in[n-1]} \frac{1}{1-\left|(\hat{H}_t)_{ii+1}\right|} \tag{27}$$

$$= \max_t \max_{i\in[n-1]} \frac{1+\left|(\hat{H}_t)_{ii+1}\right|}{1-(\hat{H}_t)^2_{ii+1}} \qquad \text{(where } (\hat{H}_t)_{ii+1} = (H_t)_{ii+1}/\sqrt{(H_t)_{ii}(H_t)_{i+1i+1}}\text{)}$$

$$\leq \max_t \max_{i\in[n-1]} \frac{2(H_t)_{ii}(H_t)_{i+1i+1}}{(H_t)_{ii}(H_t)_{i+1i+1}-(H_t)^2_{ii+1}} \qquad \text{(since } |(H_t)_{ii+1}| \leq \sqrt{(H_t)_{ii}(H_t)_{i+1i+1}}\text{)}$$

$$\leq \max_t \max_{i\in[n-1]} \frac{2(H_t)_{ii}(H_t)_{i+1i+1}}{\det\left(\begin{bmatrix}(H_t)_{ii} & (H_t)_{ii+1}\\ (H_t)_{i+1i} & (H_t)_{i+1i+1}\end{bmatrix}\right)} \tag{28}$$

Note that $\begin{bmatrix}(H_t)_{ii} & (H_t)_{ii+1}\\ (H_t)_{i+1i} & (H_t)_{i+1i+1}\end{bmatrix} \succeq \epsilon\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}$ (using line 1 in Algorithm 1), thus $\det\left(\begin{bmatrix}(H_t)_{ii} & (H_t)_{ii+1}\\ (H_t)_{i+1i} & (H_t)_{i+1i+1}\end{bmatrix}\right) \geq \det\left(\epsilon\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}\right) = \epsilon^2$. The numerator last inequality can be upperbounded by bounding $(H_t)_{ii}$ individually as follows:

$$\begin{aligned}(H_t)_{ii} &= \sum_{s=1}^{t}(g_s)_i^2/\lambda_s\\ &= \sum_{s=1}^{t}(g_s)_i^2/\lambda_s\\ &= \sum_{s=1}^{t}(g_s)_i^2/(G_\infty\sqrt{s})\\ &\leq \sum_{s=1}^{t}G_\infty^2/(G_\infty\sqrt{s})\\ &\leq \sum_{s=1}^{t}\frac{G_\infty}{\sqrt{s}}\\ &\leq 2G_\infty\sqrt{t}\end{aligned} \tag{29}$$

Substituting the above in (28) gives

$$\begin{aligned}1/(1-\beta) &\leq \max_t \frac{8G_\infty^2 t}{\hat{\epsilon}^2 G_\infty^2 T}\\ &\leq \frac{8}{\hat{\epsilon}^2}\end{aligned}$$

□

Lemma .5 (*Upperbound of T_1*)**.**

$$T_1 \leq \frac{16D_2^2 G_\infty\sqrt{T}}{\hat{\epsilon}^2\eta}, \tag{30}$$

where $D_2 = \max_{t\in[T]}\|w_t - w^*\|_2$ and $G_\infty = \max_t\|g_t\|_\infty$

Proof. Since X_T is positive definite

$$\begin{aligned} T_1 &\le \frac{\|w_1 - w^*\|^2_{X_1^{-1}}}{2\eta} \\ &= \frac{(y^{(1)})^T X_1^{-1} y^{(1)}}{2\eta} && \text{(where } y^{(1)} = w_1 - w^*) \\ &\le \frac{\|y^{(1)}\|_2^2 \|\text{diag}(H_1)\|_2}{2\eta} \cdot \frac{1+\beta}{1-\beta} && \text{(Lemma .3)} \\ &\le \frac{D_2^2(G_\infty^2/\lambda_1 + \epsilon)}{2\eta} \cdot \frac{1+\beta}{1-\beta} && \text{(line 4 in Algorithm 1)} \\ &\le \frac{8D_2^2(G_\infty^2/\lambda_1 + \epsilon)}{\hat{\epsilon}^2\eta} && \text{(Lemma .4)} \\ &\le \frac{8D_2^2(G_\infty + \hat{\epsilon}G_\infty\sqrt{T})}{\hat{\epsilon}^2\eta} && \text{(Since } \lambda_t = G_\infty\sqrt{t} \text{ and } \epsilon = \hat{\epsilon}G_\infty\sqrt{T}) \\ &\le \frac{16D_2^2 G_\infty\sqrt{T}}{\hat{\epsilon}^2\eta} && (\hat{\epsilon} < 1) \end{aligned}$$

□

Lemma .6 ($O(\sqrt{T})$ upperbound on T_3)**.**

$$T_3 = \sum_{t=1}^{T} \frac{\eta}{2} \cdot g_t^T X_t g_t \le \frac{4nG_\infty\eta}{\hat{\epsilon}^3}\sqrt{T}$$

where, $\|g_t\|_\infty \le G_\infty$ *and parameters* $\epsilon = \hat{\epsilon}G_\infty\sqrt{T}$, $\lambda_t = G_\infty\sqrt{t}$ *in Algorithm 1.*

Proof. Using Theorem 3.1, nonzero entries of X_t can be written as follows:

$$\begin{aligned} (X_t)_{ii} &= \frac{1}{H_{ii}}\left(1 + \sum_{(i,j)\in E_\mathcal{G}} \frac{H_{ij}^2}{H_{ii}H_{jj} - H_{ij}^2}\right) \\ (X_t)_{ii+1} &= -\frac{H_{ii+1}}{H_{ii}H_{i+1i+1} - H_{ii+1}^2} \end{aligned}$$

where, $E_\mathcal{G}$ denote the set of edges of the chain graph $\mathcal{G}$ in Theorem 3.1. Also, for brevity, the subscript is dropped for H_t. Let $\hat{X}_t = \sqrt{\text{diag}(H)}X_t\sqrt{\text{diag}(H)}$, then $\hat{X}_t$ can be written as

$$\begin{aligned} (\hat{X}_t)_{ii} &= \left(1 + \sum_{(i,j)\in E_\mathcal{G}} \frac{\hat{H}_{ij}^2}{1 - \hat{H}_{ij}^2}\right), \\ (\hat{X}_t)_{ii+1} &= -\frac{\hat{H}_{ii+1}}{1 - \hat{H}_{ii+1}^2}, \end{aligned}$$

where, $\hat{H}_{ij} = H_{ij}/\sqrt{H_{ii}H_{jj}}$. Note that $\hat{X}_t \preceq \|\hat{X}_t\|_2 I_n \preceq \|\hat{X}_t\|_\infty I_n$, using $\max\{|\lambda_1(\hat{X}_t)|, \ldots, |\lambda_n(\hat{X}_t)|\} \le \|\hat{X}_t\|_\infty$ (property of spectral radius). So we upperbound $\|\hat{X}_t\|_\infty = \max_{i\in[n]}\{|(\hat{X}_t)_{11}| + |(\hat{X}_t)_{12}|, \ldots, |(\hat{X}_t)_{ii-1}| + |(\hat{X}_t)_{ii}| + |(\hat{X}_t)_{ii+1}|, \ldots, |(\hat{X}_t)_{nn}| + |(\hat{X}_t)_{nn-1}|\}$ next. Individual terms $|(\hat{X}_t)_{ii-1}| + |(\hat{X}_t)_{ii}| + |(\hat{X}_t)_{ii+1}|$ can be written as follows:

$$\sum_{(i,j)\in E_{\mathcal{G}}} |(\hat{X}_t)_{ij}| = 1 + \sum_{(i,j)\in E_{\mathcal{G}}} \frac{\hat{H}_{ij}^2}{1-\hat{H}_{ij}^2} + \frac{|\hat{H}_{ij}|}{1-\hat{H}_{ij}^2}$$
$$= 1 + \sum_{(i,j)\in E_{\mathcal{G}}} \frac{|\hat{H}_{ij}|}{1-|\hat{H}_{ij}|}$$
$$\leq 2 \max_{i\in[n-1]} \frac{1}{1-|\hat{H}_{ii+1}|}$$

The last inequality is because $|\hat{H}_{ij}| \leq 1$. Thus, $\|\hat{X}_t\|_\infty \leq 2\max_{i\in[n-1]} \frac{1}{1-|\hat{H}_{ii+1}|}$. Now

$$g_t^T X_t g_t \leq g_t^T \operatorname{diag}(H_t)^{-1/2} \hat{X}_t \operatorname{diag}(H_t)^{-1/2} g_t$$
$$\leq \|\hat{X}_t\|_\infty \|\operatorname{diag}(H_t)^{-1/2} g_t\|_2^2 \qquad \left(\left\|\hat{X}_t\right\|_2 \leq \left\|\hat{X}_t\right\|_\infty\right)$$
$$\leq 2 \max_{i\in[n-1]} \frac{1}{1-|\hat{H}_{ii+1}|} g_t^T \operatorname{diag}(H_t)^{-1} g_t.$$

Using $\operatorname{diag}(H_t) \succeq \epsilon I_n$ (step 1 in Algorithm 1), where $\epsilon = \hat{\epsilon} G_\infty \sqrt{T}$ as set in Lemma A.8, gives

$$g_t^T X_t g_t \leq 2 \max_{i\in[n-1]} \frac{1}{1-|\hat{H}_{ii+1}|} \frac{\|g_t\|_2^2}{\hat{\epsilon} G_\infty \sqrt{T}}$$
$$\leq 2 \max_{i\in[n-1]} \frac{n G_\infty}{\hat{\epsilon}(1-|\hat{H}_{ii+1}|)\sqrt{T}}$$
$$\leq \frac{2nG_\infty}{\hat{\epsilon}(1-\beta)\sqrt{T}} \qquad \left(\text{where } \beta = \max_{t\in[T]} \max_{i\in[n-1]} \left|(\hat{H}_t)_{ii+1}\right|\right)$$

Summing up over t gives

$$\sum_t \frac{\eta}{2} g_t^T X_t g_t \leq \sum_t \frac{16 n G_\infty \eta}{\hat{\epsilon}^3 \sqrt{T}} \qquad \text{(Using Lemma .4)}$$
$$\leq \frac{16 n G_\infty \eta}{\hat{\epsilon}^3} \sqrt{T}$$

□

A.2.4 $\mathcal{O}(\sqrt{T})$ Regret

In this section we derive a regret upper bound with a $\mathcal{O}(T^{1/2})$ growth. For this, we upper bound T_2 as well in this section. In (21), $T_2 = \sum_{t=2}^T (w_t - w^*)^T (X_t^{-1} - X_{t-1}^{-1})(w_t - w^*)$ can be upper bounded to a $\mathcal{O}(T^{1/2})$ by upperbounding entries of $X_t^{-1} - X_{t-1}^{-1}$ individually. The following lemmas constructs a telescoping argument to bound $\left|(X_t^{-1} - X_{t-1}^{-1})_{i,j}\right|$.

Lemma .7. *Let $H, \tilde{H} \in S_n^{++}$, such that $\tilde{H} = H + gg^T/\lambda$, where $g \in \mathbb{R}^n$, then*

$$\frac{\tilde{H}_{ij}}{\sqrt{\tilde{H}_{ii}\tilde{H}_{jj}}} - \frac{H_{ij}}{\sqrt{H_{ii}H_{jj}}}$$
$$= \frac{g_i g_j}{\lambda\sqrt{\tilde{H}_{ii}\tilde{H}_{jj}}} + \frac{H_{ij}}{\sqrt{H_{ii}H_{jj}}}\left(\sqrt{\frac{H_{ii}H_{jj}}{\tilde{H}_{ii}\tilde{H}_{jj}}} - 1\right) = \theta_{ij}$$

Proof.

$$\begin{aligned}
&\frac{\tilde{H}_{ij}}{\sqrt{\tilde{H}_{ii}\tilde{H}_{jj}}} - \frac{H_{ij}}{\sqrt{H_{ii}H_{jj}}} \\
&= \frac{1}{\sqrt{H_{ii}H_{jj}}}(\tilde{H}_{ij}\frac{\sqrt{H_{ii}H_{jj}}}{\sqrt{\tilde{H}_{ii}\tilde{H}_{jj}}} - H_{ij}) \\
&= \frac{1}{\sqrt{H_{ii}H_{jj}}}\left(g_i g_j \frac{\sqrt{H_{ii}H_{jj}}}{\sqrt{\tilde{H}_{ii}\tilde{H}_{jj}}} + H_{ij}\left(\frac{\sqrt{H_{ii}H_{jj}}}{\sqrt{\tilde{H}_{ii}\tilde{H}_{jj}}} - 1\right)\right)
\end{aligned}$$

□

The following Lemma bounds the change in the inverse of preconditioner Y^{-1}, when there is a rank one perturbation to $H \succ 0$ in following LogDet problem (11) :

$$\begin{aligned}
Y &= \underset{X \in S_n(\mathcal{G})^{++}}{\arg\min} \; -\log\det(X) + \mathrm{Tr}(XH) \\
&= \underset{X \in S_n(\mathcal{G})^{++}}{\arg\min} \; \mathrm{D}_{\ell\mathrm{d}}(X, H)
\end{aligned}$$

Lemma .8 (*Rank 1 perturbation of LogDet problem (11)*)**.** *Let $H, \tilde{H} \in S_n^{++}$, such that $\tilde{H} = H + gg^T/\lambda$, where $g \in \mathbb{R}^n$. Also, $\tilde{Y} = \arg\min_{X \in S_n(\mathcal{G})^{++}} \mathrm{D}_{\ell\mathrm{d}}(X, \tilde{H})$ and $Y = \arg\min_{X \in S_n(\mathcal{G})^{++}} \mathrm{D}_{\ell\mathrm{d}}(X, H)$, where $\mathcal{G}$ is a chain graph, then*

$$\left|(\tilde{Y}^{-1} - Y^{-1})_{ii+k}\right| \le G_\infty^2 \kappa(k\beta + k + 2)\beta^{k-1}/\lambda,$$

where $i, i+k \le n$, $G_\infty = \|g\|_\infty$ and $\max_{i,j} |H_{ij}|/\sqrt{H_{ii}H_{jj}} \le \beta < 1$. Let $\kappa(\mathrm{diag}(H)) :=$ condition number of the diagonal part of H, then $\kappa := \max(\kappa(\mathrm{diag}(H)), \kappa(\mathrm{diag}(\tilde{H})))$.

Proof. Using Lemma .2 will give the following:

$$\begin{aligned}
&\left|(\tilde{Y}^{-1} - Y^{-1})_{ii+k}\right| \\
&= \left|\frac{\tilde{H}_{ii+1}\ldots\tilde{H}_{i+k-1i+k}}{\tilde{H}_{i+1i+1}\ldots\tilde{H}_{i+k-1i+k-1}} - \frac{H_{ii+1}\ldots H_{i+k-1i+k}}{H_{i+1i+1}\ldots H_{i+k-1i+k-1}}\right| \\
&= \Big|\sqrt{\tilde{H}_{ii}}\tilde{N}_{ii+1}\ldots\tilde{N}_{i+k-1i+k}\sqrt{\tilde{H}_{i+ki+k}} \\
&- \sqrt{H_{ii}}N_{ii+1}\ldots N_{i+k-1i+k}\sqrt{H_{i+ki+k}}\Big| \\
&= \sqrt{\tilde{H}_{ii}\tilde{H}_{i+ki+k}}\left|\tilde{N}_{ii+1}\ldots\tilde{N}_{i+k-1i+k} - N_{ii+1}\ldots N_{i+k-1i+k}\sqrt{H_{ii}H_{i+ki+k}/\tilde{H}_{ii}\tilde{H}_{i+ki+k}}\right|
\end{aligned}$$

where $N_{ij} = H_{ij}/\sqrt{H_{ii}H_{jj}} < 1$ (Since determinants of 2x2 submatrices of H are positive). Expanding $\tilde{N}_{ii+1} = N_{ii+1} + \theta_{ii+1}$ (from Lemma .7), subsequently $\tilde{N}_{ii+2} = N_{ii+2} + \theta_{ii+2}$ and so on will give

$$\left|\tilde{N}_{ii+1}\ldots\tilde{N}_{i+k-1i+k} - N_{ii+1}\ldots N_{i+k-1i+k}\sqrt{\frac{H_{ii}H_{i+ki+k}}{\tilde{H}_{ii}\tilde{H}_{i+ki+k}}}\right| =$$

$$\left|\theta_{ii+1}\tilde{N}_{i+1i+2}\ldots\tilde{N}_{i+k-1i+k} + N_{ii+1}\left(\tilde{N}_{i+1i+2}\ldots\tilde{N}_{i+k-1i+k} - N_{i+1i+2}\ldots N_{i+k-1i+k}\sqrt{\frac{H_{ii}H_{i+ki+k}}{\tilde{H}_{ii}\tilde{H}_{i+ki+k}}}\right)\right|$$

$$
\begin{aligned}
&= |\theta_{ii+1}\tilde{N}_{i+1i+2}\dots\tilde{N}_{i+k-1i+k} + N_{ii+1}\theta_{i+1i+2}\tilde{N}_{ii+3}\dots\tilde{N}_{i+k-1i+k} + \dots + N_{ii+1}\dots N_{ii+k-1}\theta_{i+k-1i+k} \\
&+ N_{ii+1}\dots N_{ii+k}\left(1 - \sqrt{\frac{H_{ii}H_{i+ki+k}}{\tilde{H}_{ii}\tilde{H}_{i+ki+k}}}\right)| \\
&\leq (\sum_{l=0}^{k-1}|\theta_{i+li+l+1}|)\beta^{k-1} + \beta^{k-1}\left|1 - \sqrt{\frac{H_{ii}H_{i+ki+k}}{\tilde{H}_{ii}\tilde{H}_{i+ki+k}}}\right|, \\
&\implies \left|(\tilde{Y}^{-1} - Y^{-1})_{ii+k}\right| \leq \sqrt{\tilde{H}_{ii}\tilde{H}_{i+ki+k}} \cdot \left((\sum_{l=0}^{k-1}|\theta_{i+li+l+1}|)\beta^{k-1} + \beta^{k-1}\left|1 - \sqrt{\frac{H_{ii}H_{i+ki+k}}{\tilde{H}_{ii}\tilde{H}_{i+ki+k}}}\right|\right)
\end{aligned}
$$

where $\max_{i,j}|N_{i,j}|,\ \max_{i,j}|\tilde{N}_{i,j}| \leq \beta < 1$. Expanding $\theta_{i+li+l+1}$ from Lemma .7 in the term $|\theta_{i+li+l+1}|\sqrt{\tilde{H}_{ii}\tilde{H}_{i+ki+k}}$ will give:

$$
\begin{aligned}
&|\theta_{i+li+l+1}|\sqrt{\tilde{H}_{ii}\tilde{H}_{i+ki+k}} \\
&= \left|\sqrt{\tilde{H}_{ii}\tilde{H}_{i+ki+k}}\frac{g_{i+l}g_{i+l+1}}{\lambda\sqrt{\tilde{H}_{i+li+l}\tilde{H}_{i+l+1i+l+1}}} + \sqrt{\tilde{H}_{ii}\tilde{H}_{i+ki+k}}N_{i+li+l+1}\left(\sqrt{\frac{H_{i+li+l}H_{i+l+1i+l+1}}{\tilde{H}_{i+li+l}\tilde{H}_{i+l+1i+l+1}}} - 1\right)\right| \\
&\leq \left|\sqrt{\tilde{H}_{ii}\tilde{H}_{i+ki+k}}\frac{g_{i+l}g_{i+l+1}}{\lambda\sqrt{\tilde{H}_{i+li+l}\tilde{H}_{i+l+1i+l+1}}}\right| + \left|\sqrt{\tilde{H}_{ii}\tilde{H}_{i+ki+k}}N_{i+li+l+1}\left(1 - \sqrt{\frac{H_{i+li+l}H_{i+l+1i+l+1}}{\tilde{H}_{i+li+l}\tilde{H}_{i+l+1i+l+1}}}\right)\right|
\end{aligned}
$$

Since $H_{i+li+l}H_{i+l+1i+l+1} \leq \tilde{H}_{i+li+l}\tilde{H}_{i+l+1i+l+1}$,

$$
\begin{aligned}
1 - \sqrt{\frac{H_{i+li+l}H_{i+l+1i+l+1}}{\tilde{H}_{i+li+l}\tilde{H}_{i+l+1i+l+1}}} &\leq \max\left(1 - \frac{H_{i+li+l}}{\tilde{H}_{i+li+l}}, 1 - \frac{H_{i+l+1i+l+1}}{\tilde{H}_{i+l+1i+l+1}}\right) \\
&\leq \max\left(\frac{g_{i+l}^2}{\lambda\tilde{H}_{i+li+l}}, \frac{g_{i+l+1}^2}{\lambda\tilde{H}_{i+l+1i+l+1}}\right)
\end{aligned}
$$

Using the above, $H_{i,i}/H_{j,j} \leq \kappa$, and $|g_i| \leq G_\infty, \forall i,j \in [n]$, gives

$$
\begin{aligned}
\sqrt{\tilde{H}_{ii}\tilde{H}_{i+ki+k}}|\theta_{i+li+l+1}| &\leq G_\infty^2\kappa/\lambda + \beta G_\infty^2\kappa/\lambda \\
&\leq G_\infty^2\kappa(1+\beta)/\lambda
\end{aligned}
$$

Thus the following part of $\left|\left(\tilde{Y}^{-1} - Y^{-1}\right)_{ii+k}\right|$ can be upperbounded:

$$
\sqrt{\tilde{H}_{ii}\tilde{H}_{i+ki+k}}\left((\sum_{l=0}^{k-1}|\theta_{i+li+l+1}|)\beta^{k-1}\right) \leq G_\infty^2\kappa(1+\beta)k\beta^{k-1}/\lambda
$$

Also, $\sqrt{\tilde{H}_{ii}\tilde{H}_{i+ki+k}}\beta^{k-1}\left|1 - \sqrt{\frac{H_{ii}H_{i+ki+k}}{\tilde{H}_{ii}\tilde{H}_{i+ki+k}}}\right| \leq \beta^{k-1}\kappa G_\infty^2/\lambda$, so

$$
\left|\left(\tilde{Y}^{-1} - Y^{-1}\right)_{ii+k}\right| \leq G_\infty^2\kappa(k\beta + k + 2)\beta^{k-1}/\lambda
$$

□

Lemma .9 **($\mathcal{O}(\sqrt{T})$ upper bound of T_2).** *Given that $\kappa(\text{diag}(H_t)) \leq \kappa$, $\|w_t - w^*\|_2 \leq D_2$, $\max_{i,j}|(H_t)_{ij}|/\sqrt{(H_t)_{ii}(H_t)_{jj}} \leq \beta < 1$, $\forall t \in [T]$ in Algorithm 1, then T_2 in Appendix A.2.1 can be bounded as follows:*

$$
T_2 \leq \frac{2048\sqrt{T}}{\eta\hat{\epsilon}^5}(G_\infty D_2^2)
$$

where $\lambda_t = G_\infty\sqrt{t}$, and $\epsilon = \hat{\epsilon}G_\infty\sqrt{T}$ in Algorithm 1, and $\hat{\epsilon} \leq 1$ is a constant.

Proof. Note that $T_2 = \frac{1}{2\eta} \cdot \sum_{t=1}^{T-1}(w_{t+1} - w^*)^T(X_{t+1}^{-1} - X_t^{-1})(w_{t+1} - w^*) \leq \sum_{t=1}^{T-1} D_2^2 \left\|(X_{t+1}^{-1} - X_t^{-1})\right\|_2 /(2\eta)$. Using $\|A\|_2 = \rho(A) \leq \|A\|_\infty$ for symmetric matrices A, we get

$$\begin{aligned}
\left\|X_{t+1}^{-1} - X_t^{-1}\right\|_2 &\leq \|X_{t+1}^{-1} - X_t^{-1}\|_\infty \\
&= \max_i(\sum_j \left|(X_{t+1}^{-1} - X_t^{-1})_{ij}\right|) \\
&\leq 16\frac{G_\infty \kappa}{\sqrt{t}(1-\beta)^2} && \text{(Lemma .8)} \\
&\leq 1024 \cdot \frac{G_\infty \kappa}{\sqrt{t}\hat{\epsilon}^4}
\end{aligned}$$

Now using $\kappa \leq 2/\hat{\epsilon}$ (using Equation (29) and $(H_t)_{ii} > \hat{\epsilon}$) and summing up terms in T_2 using the above will give the result. ☐

Putting together T_1, T_2 and T_3 from Lemma .5, Lemma .9 and Lemma .6 respectively, when ϵ, λ_t are defined as in Lemma .9:

$$\begin{aligned}
T_1 \leq& \frac{16 D_2^2 G_\infty \sqrt{T}}{\hat{\epsilon}^2 \eta}, \\
T_2 \leq& \frac{2048\sqrt{T}}{\eta\hat{\epsilon}^5}(G_\infty D_2^2) && (31) \\
T_3 \leq& \frac{4nG_\infty \eta}{\hat{\epsilon}^3}\sqrt{T} && (32)
\end{aligned}$$

Setting $\eta = \frac{D_2}{\hat{\epsilon}\sqrt{n}}$

$$R_T \leq T_1 + T_2 + T_3 \leq O(\sqrt{n} G_\infty D_2 \sqrt{T})$$

A.2.5 Non-convex guarantees

Minimizing smooth non-convex functions f is a complex yet interesting problem. In Agarwal et al. [1], this problem is reduced to an online convex optimization, where a sequence of objectives $f_t(w) = f(w) + c\|w - w_t\|_2^2$ are minimized. Using this approach Agarwal et al. [1] established convergence guarantees to reach a stationary point via regret minimization. Thus non-convex guarantees can be obtained from regret guarantees and is our main focus in the paper.

A.3 Numerical stability

In this section we conduct perturbation analysis to derive an end-to-end componentwise condition number (pg. 135, problem 7.11 in [26]) upper bound of the tridiagonal explicit solution in Theorem 3.1. In addition to this, we devise Algorithm 3 to reduce this condition number upper bound for the tridiagonal sparsity structure, and be robust to H_t which don't follow the non-degeneracy condition: any principle submatrix of H_t corresponding to a complete subgraph of $\mathcal{G}$.

Theorem .10 (Condition number of tridiagonal LogDet subproblem (11))**.** *Let $H \in S_n^{++}$ be such that $H_{ii} = 1$ for $i \in [n]$. Let ΔH be a symmetric perturbation such that $\Delta H_{ii} = 0$ for $i \in [n]$, and $H + \Delta H \in S_n^{++}$. Let $P_\mathcal{G}(H)$ be the input to 11, where $\mathcal{G}$ is a chain graph, then*

$$\kappa_\infty^{\ell d} \leq \max_{i\in[n-1]} 2/(1-\beta_i^2) = \hat{\kappa}_\infty^{\ell d}, \tag{33}$$

where, $\beta_i = H_{ii+1}$,$\kappa_\infty^{\ell d}$:= componentwise condition number of (11) for perturbation ΔH.

The tridiagonal LogDet problem with inputs H as mentioned in Theorem .10, has high condition number when $1 - \beta_i^2 = H_{ii} - H_{ii+1}^2/H_{i+1i+1}$ are low and as a result the preconditioner X_t in

SONew (Algorithm 1) has high componentwise relative errors. We develop Algorithm 3 to be robust to degenerate inputs H, given that $H_{ii} > 0$. It finds a subgraph $\tilde{\mathcal{G}}$ of $\mathcal{G}$ for which non-degeneracy conditions in Theorem 3.2 is satisfied and (14) is well-defined. This is done by removing edges which causes inverse $H^{-1}_{I_j I_j}$ to be singular or $(H_{jj} - H^T_{I_j j} H^{-1}_{I_j I_j} H_{I_j j})$ to be low. In the following theorem we also show that the condition number upper bound in Theorem .10 reduces in tridiagonal case. To test the robustness of this method we conducted an ablation study in Table 5, in an Autoencoder benchmark (from Section 5) in bfloat16 where we demonstrate noticeable improvement in performance when Algorithm 3 is used.

Theorem .11 (Numerically stable algorithm)**.** *Algorithm 3 finds a subgraph $\tilde{\mathcal{G}}$ of $\mathcal{G}$, such that explicit solution for $\tilde{\mathcal{G}}$ in (14) is well-defined. Furthermore, when $\mathcal{G}$ is a tridiagonal/chain graph, the component-wise condition number upper bound in (33) is reduced upon using Algorithm 3, $\hat{\kappa}^{\tilde{\mathcal{G}}}_{\ell d} < \hat{\kappa}^{\mathcal{G}}_{\ell d}$, where $\hat{\kappa}^{\tilde{\mathcal{G}}}_{\ell d}$, $\hat{\kappa}^{\mathcal{G}}_{\ell d}$ are defined as in Theorem .10 for graphs $\tilde{\mathcal{G}}$ and $\mathcal{G}$ respectively.*

The proofs for Theorems .10 and .11 are given in the following subsections.

Algorithm 3 Numerically stable banded LogDet solution

Input: $\mathcal{G}-$ tridiagonal or banded graph, $H-$ symmetric matrix in $\mathbb{R}^{n\times n}$ with sparsity structure $\mathcal{G}$ and $H_{ii} > 0$, $\gamma-$ tolerance parameter for low schur complements.
Output: Finds subgraph $\tilde{\mathcal{G}}$ of $\mathcal{G}$ without any degenerate cases from Lemma .13 and finds preconditioner $\hat{X}$ corresponding to the subgraph
Let $E_i = \{(i,j) : (i,j) \in E_{\mathcal{G}}\}$ be edges from vertex i to its neighbours in graph $\mathcal{G}$.
Let $V_i^+ = \{j : i < j, (i,j) \in E_{\mathcal{G}}\}$ and $V_i^- = \{j : i > j, (i,j) \in E_{\mathcal{G}}\}$, denote positive and negative neighbourhood of vertex i.
Let $K = \left\{ i : H_{ii} - H^T_{I_i i} H^{-1}_{I_i I_i} H_{I_i i} \text{ is undefined or } \leq \gamma \right\}$
Consider a new subgraph $\tilde{\mathcal{G}}$ with edges $E_{\tilde{\mathcal{G}}} = E_{\mathcal{G}} \setminus (\bigcup_{i\in K} E_i \cup (V_i^+ \times V_i^-))$
return $\hat{X} :=$ SPARSIFIED_INVERSE $(\tilde{H}_t, \tilde{\mathcal{G}})$, where $\tilde{H}_t = P_{\tilde{\mathcal{G}}}(H_t)$

A.3.1 Condition number analysis

Theorem .12 (*Full version of Theorem .10*)**.** *Let $H \in S_n^{++}$ such that $H_{ii} = 1$, for $i \in [n]$ and a symmetric perturbation ΔH such that $\Delta H_{ii} = 0$, for $i \in [n]$ and $H + \Delta H \succ 0$. Let $\hat{X} = \arg\min_{X \in S_n(\mathcal{G})^{++}} \mathrm{D}_{\ell \mathrm{d}}\left(X, H^{-1}\right)$ and $\hat{X} + \Delta \hat{X} = \arg\min_{X\in S_n(\mathcal{G})^{++}} \mathrm{D}_{\ell \mathrm{d}}\left(X, (H+\Delta H)^{-1}\right)$, here $\mathcal{G} :=$ chain/tridiagonal sparsity graph and $S_n(\mathcal{G})^{++}$ denotes positive definite matrices which follows the sparsity pattern $\mathcal{G}$.*

$$\begin{aligned}\kappa_{\ell d} &= \lim_{\epsilon\to 0} \sup \left\{ \frac{\left|\Delta \hat{X}_{ij}\right|}{\epsilon \left|\hat{X}_{ij}\right|} : |\Delta H_{k,l}| \leq |\epsilon H_{k,l}|, (k,l) \in E_{\mathcal{G}} \right\} \\ &\leq \max_{i\in[n-1]} 1/(1-\beta_i^2)\end{aligned}$$

where, $\kappa_{\ell d} :=$ condition number of the LogDet subproblem, $\kappa_2(.) :=$ condition number of a matrix in ℓ_2 norm, $\beta_i = H_{ii+1}/\sqrt{H_{ii}H_{i+1i+1}}$

Proof. Consider the offdiagonals for which $(\hat{X} + \Delta\hat{X})_{ii+1} = -H_{ii+1}/(1 - H^2_{ii+1}) = f(H_{ii+1})$,where $f(x) = -x/(1-x^2)$. Let $y = f(x)$, $\hat{y} = f(x + \Delta x)$ and $|\Delta x/x| \leq \epsilon$ then using Taylor series

$$\begin{aligned}\left|\frac{(\hat{y}-y)}{y}\right| &= \left|\frac{xf'(x)}{f(x)}\right| \left|\frac{\Delta x}{x}\right| + O((\Delta x)^2) \\ \implies \lim_{\epsilon\to 0}\left|\frac{(\hat{y}-y)}{\epsilon y}\right| &\leq \frac{xf'(x)}{f(x)}\end{aligned}$$

Using the above inequality, with $x := H_{ii+1}$ and $y := \hat{X}_{ii+1}$,

$$\lim_{\epsilon\to 0}\left|\frac{\Delta\hat{X}_{ii+1}}{\epsilon\hat{X}_{ii+1}}\right| \leq \frac{1+H_{ii+1}^2}{1-H_{ii+1}^2} \leq \frac{2}{1-H_{ii+1}^2} \tag{34}$$

Let $g(x) = x^2/(1-x^2)$, let $y_1 = g(w_1)$, $y_2 = g(x_2)$, $\hat{y}_1 = g(w_1+\Delta x)$, $\hat{y}_2 = g(x_2+\Delta x)$. Using Taylor series

$$\left|\frac{(\hat{y}_1-y_1)}{y_1}\right| = \left|\frac{x_1 f'(x_1)}{f(x_1)}\right|\left|\frac{\Delta x_1}{x_1}\right| + O((\Delta x_1)^2)$$
$$\left|\frac{(\hat{y}_2-y_2)}{y_2}\right| = \left|\frac{x_2 f'(x_2)}{f(x_2)}\right|\left|\frac{\Delta x_2}{x_2}\right| + O((\Delta x_2)^2)$$
$$\implies \lim_{\epsilon\to 0}\frac{\Delta y_1+\Delta y_2}{\epsilon(1+y_1+y_2)} \leq \max\left(\frac{2}{1-x_1^2}, \frac{2}{1-x_2^2}\right)$$

Putting $x_1 := H_{ii+1}$, $x_2 := H_{ii-1}$ and analyzing $y_1 := H_{ii+1}^2/(1-H_{ii+1}^2)$ and $y_2 := H_{ii-1}^2/(1-H_{ii-1}^2)$ will result in the following

$$\lim_{\epsilon\to 0}\left|\frac{\Delta\hat{X}_{ii}}{\hat{X}_{ii}}\right| \leq \max\left(\frac{2}{1-H_{ii+1}^2}, \frac{2}{1-H_{ii-1}^2}\right) \tag{35}$$

Since $\hat{X}_{ii} = 1 + H_{ii+1}^2/(1-H_{ii+1}^2) + H_{ii-1}^2/(1-H_{ii-1}^2)$. Putting together Equation (35) and Equation (34), the theorem is proved. □

A.3.2 Degenerate H_t

In SONew (1), the $H_t = P_{\mathcal{G}}(\sum_{s=1}^{t} g_s g_s^T/\lambda_t)$ generated in line 4 could be such that the matrix $\sum_{s=1}^{t} g_s g_s^T/\lambda_t$ need not be positive definite and so the schur complements $H_{ii} - H_{ii+1}^2/H_{i+1i+1}$ can be zero, giving an infinite condition number $\kappa_\infty^{\ell d}$ by Theorem .10. The following lemma describes such cases in detail for a more general banded sparsity structure case.

Lemma .13 (Degenerate inputs to banded LogDet subproblem)**.** *Let $H = P_{\mathcal{G}}(GG^T)$, when $\epsilon = 0$ in Algorithm 1, where $G \in \mathbb{R}^{n\times T}$ and let $g_{1:T}^{(i)}$ be i^{th} row of G, which is gradients of parameter i for T rounds, then $H_{ij} = \left\langle g_{1:T}^{(i)}, g_{1:T}^{(j)}\right\rangle$.*

- *Case 1: For tridiagonal sparsity structure $\mathcal{G}$: if $g_{1:T}^{(j)} = g_{1:T}^{(j+1)}$, then $H_{jj} - H_{jj+1}^2/H_{j+1j+1} = 0$.*
- *Case 2: For $b > 1$ in (14): If $\operatorname{rank}(H_{J_jJ_j}) = \operatorname{rank}(H_{I_jI_j}) = b$, then $(H_{jj} - H_{I_jj}^T H_{I_jI_j}^{-1} H_{I_jj}) = 0$ and $D_{jj} = \infty$. If $\operatorname{rank}(H_{I_jI_j}) < b$ then the inverse $H_{I_jI_j}^{-1}$ doesn't exist and D_{jj} is not well-defined.*

Proof. For $b = 1$, if $g_{1:T}^{(j)} = g_{1:T}^{(j+1)}$, then $H_{jj+1} = H_{jj} = H_{j+1j+1} = \left\|g_{1:T}^{(j)}\right\|_2^2$, thus $H_{jj} - H_{jj+1}^2/H_{j+1j+1} = 0$.
For $b > 1$, since $H_{I_jI_j}$, using Guttman rank additivity formula, $\operatorname{rank}(H_{jj} - H_{jj+1}^2/H_{j+1j+1}) = \operatorname{rank}(H_{J_jJ_j}) - rank(H_{I_jI_j}) = 0$, thus $H_{jj} - H_{jj+1}^2/H_{j+1j+1} = 0$.
Furthermore, if $\operatorname{rank}(H) \leq b$, then all $b+1 \times b+1$ principal submatrices of H have rank b, thus $\forall j$, $H_{J_jJ_j}$ have a rank b, thus D_{jj} for all j are undefined.

□

If $GG^T = \sum_{i=1}^{T} g_i g_i$ is a singular matrix, then solution to the LogDet problem might not be well-defined as shown in Lemma .13. For instance, Case 1 can occur when preconditioning the input layer of an image-based DNN with flattened image inputs, where j^{th} and $(j+1)^{th}$ pixel can be highly correlated throughout the dataset. Case 2 can occur in the first b iterations in Algorithm 1 when the rank of submatrices $\operatorname{rank}(H_{I_jI_j}) < b$ and $\epsilon = 0$.

Table 3: **float32 experiments on Autoencoder benchmark using different band sizes.** Band size 0 corresponds to diag-SONew and 1 corresponds to tridiag-SONew. We see the training loss getting better as we increase band size

Band size	0 (diag-SONew)	1 (tridiag-SONew)	4	10
Train CE loss	53.025	51.723	51.357	51.226

A.3.3 Numerically Stable SONew proof

Proof of Theorem .11

Let $I_i = \{j : i < j, (i,j) \in E_{\mathcal{G}}\}$ and $I'_i = \{j : i < j, (i,j) \in E_{\tilde{\mathcal{G}}}\}$ Let $K = \{i : H_{ii} - H_{I_i i}^T H_{I_i I_i}^{-1} H_{I_i i}$ is undefined or $0, i \in [n]\}$ denote vertices which are getting removed by the algorithm, then for the new graph $\tilde{\mathcal{G}}$, $D_{ii} = 1/H_{ii}, \forall i \in K$ since $H_{ii} > 0$.
Let $\bar{K} = \{i : H_{ii} - H_{I_i i}^T H_{I_i I_i}^{-1} H_{I_i i} > 0, i \in [n]\}$. Let for some $j \in \bar{K}$, if

$$l = \arg\min \{i : j < i, i \in K \cap I_j\},$$

denotes the nearest connected vertex higher than j for which D_{ll} is undefined or zero, then according to the definition $E_{\tilde{\mathcal{G}}}$ in Algorithm 3, $I'_j = \{j+1, \ldots l-1\} \subset I_j$, since D_{jj} is well-defined, $H_{I_j I_j}$ is invertible, which makes it a positive definite matrix (since H is PSD). Since $H_{jj} - H_{I_j j}^T H_{I_j I_j}^{-1} H_{I_j j} > 0$, using Guttman rank additivity formula $H_{J_j J_j} \succ 0$, where $J_j = I_j \cup j$. Since $H_{J'_j J'_j}$ is a submatrix of $H_{J_j J_j}$, it is positive definite and hence its schur complement $H_{jj} - H_{I'_j j}^T H_{I'_j I'_j}^{-1} H_{I'_j j} > 0$. Thus for all $j \in [n]$, the corresponding D_{jj}'s are well-defined in the new graph $\tilde{\mathcal{G}}$.

Note that $\kappa_{\ell d}^{\tilde{\mathcal{G}}} = \max_{i \in [n-1]} 1/(1-\beta_i^2) < \max_{i \in \bar{K}} 1/(1-\beta_i^2) = \kappa_{\ell d}^{\mathcal{G}}$, for tridiagonal graph, where $\beta_i = H_{ii+1}$, in the case where $H_{ii} = 1$. This is because the $\arg\max_{i \in [n-1]} 1/(1-\beta_i^2) \in K$.

A.4 Additional Experiments, ablations, and details

A.4.1 Ablations

Effect of band size in banded-SONew Increasing band size in banded-SONew captures more correlation between parameters, hence should expectedly lead to better preconditioners. We confirm this through experiments on the Autoencoder benchmark where we take band size = 0 (diag-SONew), 1 (tridiag-SONew), 4, and 10 in Table 3.

Effect of mini-batch size To find the effect of mini-batch size, in Table 4, We empirically compare SONew with state of the art first-order methods such as Adam and RMSProp, and second-order method Shampoo. We see that SONew performance doesn't deteriorate much when using smaller or larger batch size. First order methods on the other hand suffer significantly. We also notice that Shampoo doesn't perform better than SONew in these regimes.

Table 4: **Comparison on Autoencoder with different batch-sizes**

Baseline\Batch size	**100**	**1000**	**5000**	**10000**
RMSProp	55.61	53.33	58.69	64.91
Adam	55.67	54.39	58.93	65.37
Shampoo(20)	53.91	50.70	53.52	54.90
tds	53.84	51.72	54.24	55.87
bds-4	53.52	51.35	53.03	54.89

Effect of Numerical Stability Algorithm 3 On tridiag-SONew and banded-4-SONew, we observe that using Algorithm 3 improves training loss. We present in Table 5 results where we observed significant performance improvements.

Table 5: **bfloat16 experiments on Autoencoder benchmark with and without Algorithm 3.** We observe improvement in training loss when using Algorithm 3

Optimizer	Train CE loss - without Algorithm 3	Train CE loss - with Algorithm 3
tridiag-SONew	53.150	51.936
band-4-SONew	51.950	51.84

A.4.2 Hyperparaeter search space

We provide the hyperparamter search space for experiments presented in Section 5. We search over $2k$ hyperparameters for each Autoencoder experiment using a Bayesian Optimization package. The search ranges are: first order momentum term $\beta_1 \in [1e-1, 0.999]$, second order momentum term $\beta_2 \in [1e-1, 0.999]$, learning rate $\in [1e-7, 1e-1]$, $\epsilon \in [1e-10, 1e-1]$. We give the optimal hyperparameter value for each experiment in Table 11. For VIT and GraphNetwork benchmark, we search $\beta_1, \beta_2 \in [0.1, 0.999]$, $lr \in [1e-5, 1e-1]$, $\epsilon \in [1e-9, 1e-4]$, weight decay $\in [1e-5, 1.0]$, learning rate warmup $\in [2\%, 5\%, 10\%]*$total_train_steps, dropout$\in [00, 0.1]$, label smoothing over $\{0.0, 0.1, 0.2\}$. We use cosine learning rate schedule. Batch size was kept = 1024, and 512 for Vision Transformer, and GraphNetwork respectively. We sweep over 200 hyperparameters in the search space for all the optimizers.
For rfdSON [36], there's no ϵ hyperparameter. In addition to the remaining hyperparameters, we tune $\alpha \in \{1e-5, 1.0\}$ (plays similar role as ϵ) and $\mu_t \in [1e-5, 0.1]$.
For LLM [44] benchmark, we only tune the learning rate $\in [1e-2, 1e-3, 1e-4]$ while keeping the rest of the hyperparams as constant. This is due to the high cost of running experiments hence we only tune the most important hyperparameter. For Adafactor [43], we use factored=False, decay method=adam, $\beta_1 = 0.9$, weight decay=$1e-3$, decay factor=0.99, and gradient clipping=1.0.

A.4.3 Additional Experiments

VIT and GraphNetwork Benchmarks: In Figure 5 we plot the training loss curves of runs corresponding to the best validation runs in Figure 1. Furthermore, from an optimization point of view, we plot the best train loss runs in Figure 6 got by searching over 200 hyperparameters. We find that tridiag-SONew is 9% and 80% relatively better in ViT and GraphNetwork benchmark respectively (Figure 6), compared to Adam (the next best baseline).

Autoencoder float32 and bfloat16 experiments: We provide curves of all the baselines and SONew in Figure 4(a) and the corresponding numbers in Table 6 for float32 experiments.

To test numerical stability of SONew and compare it with other algorithm in low precision regime, we also conduct bfloat16 experiments on the Autoencoder benchmark (Table 7). We notice that SONew undergoes the least degradation. Tridiagonal-sparsity SONew CE loss increases by only 0.21 absolute difference (from 51.72 in float32 (6) to 51.93), whereas Shampoo and Adam incur 0.70 loss increase. It's worthwhile to note that SONew performs better than all first order methods while taking similar time and linear memory, whereas while Shampoo performs marginally better, it is $22\times$ slower than tridiagonal-SONew. The corresponding loss curves are given in Figure 4(b).
Note: In the main paper, our reported numbers for rfdSON on Autoencoder benchmark in Table 1 for float32 experiments are erraneuous. Please consider the numbers provided in Table 6 and the corresponding curve in Figure 4(a). Note that there's no qualitiative change in the results and none of the claims made in the paper are affected. SONew is still significantly better than rfdSON. We also meticulously checked all other experiments, and they do not have any errors.

A.4.4 Convex experiments

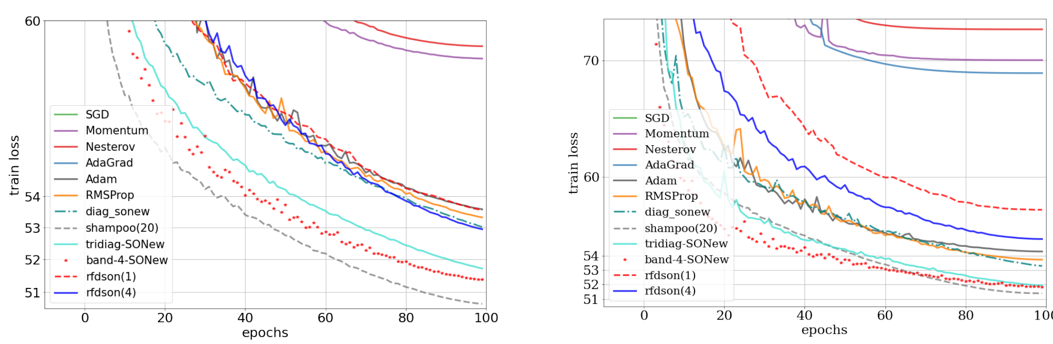
As our regret bound applies to convex optimization, we compare SONew to rfdSON [36], another recent memory-efficient second-order Newton method. We follow [36] for the experiment setup - each dataset is split randomly in 70%/30% train and test set. Mean squared loss is used. For tridiag-SONew, we use a total of $2*d$ space for d parameters. Hence, for fair comparison we show rfdSON with $m = 2$. Since the code isn't open sourced, we implemented it ourselves. In order to show reproducibility with respect to the reported numbers in [36], we include results with $m = 5$ as well. We see in the Table 8 that tridiag-SONew consitently matches or outperforms rfdSON across all

Table 6: **float32 experiments on Autoencoder benchmark.** We observe that diag-SONew performs the best among all first order methods while taking similar time. tridiag and band-4 perform significantly better than first order methods while requiring similar linear space and time. Shampoo performs best but takes $\mathcal{O}(d_1^3 + d_2^3)$ time for computing preconditioner of a linear layer of size $d_1 \times d_2$, whereas our methods take $\mathcal{O}(d_1 d_2)$ time, as mentioned in Section 5.1. rfdSON takes similar space as SONew but performs considerably worse.

Optimizer	**First Order Methods**						
	SGD	**Nesterov**	**Adagrad**	**Momentum**	**RMSProp**	**Adam**	**diag-SONew**
Train CE loss	67.654	59.087	54.393	58.651	53.330	53.591	53.025
Time(s)	62	102	62	67	62	62	63
Optimizer	**Second Order Methods**						
	Shampoo(20)	**rfdSON(1)**	**rfdSON(4)**	**tridiag-SONew**	**band-4-SONew**		
Train CE loss	50.702	53.56	52.97	51.723	51.357		
Time(s)	371	85	300	70	260		

Table 7: **bfloat16 experiments on Autoencoder benchmark** to test the numerical stability of SONew and robustness of Algorithm 3. We notice that diag-SONew degrades only marginally (0.26 absolute difference) compared to float32 performance. tridiag-SONew and band-4-SONew holds similar observations as well. Shampoo performs the best but has a considerable drop (0.70) in performance compared to float32 due to using matrix inverse, and is slower due to its cubic time complexity for computing preconditioners. Shampoo implementation uses 16-bit quantization to make it work in 16-bit setting, leading to further slowdown. Hence the running time in bfloat16 is even higher than in float32.

Optimizer	**First Order Methods**						
	SGD	**Nesterov**	**Adagrad**	**Momentum**	**RMSProp**	**Adam**	**diag-SONew**
Train CE loss	80.454	72.975	68.854	70.053	53.743	54.328	53.29
Train time(s)	36	43	37	36	37	38	44
Optimizer	**Second Order Methods**						
	Shampoo(20)	**rfdSON(1)**	**rfdSON(4)**	**tridiag-SONew**	**band-4-SONew**		
Train CE loss	51.401	57.42	55.53	51.937	51.84		
Train time(s)	1245	80	284	55	230		



(a) float32 - autoencoder

(b) bfloat16 - autoencoder

Figure 4: Training curves of all the baselines for Autoencoder benchmar (a) float32 training (b) bfloat16 training

3 benchmarks. Each experiment was run for 20 epochs and we report the best model's performance on test set.

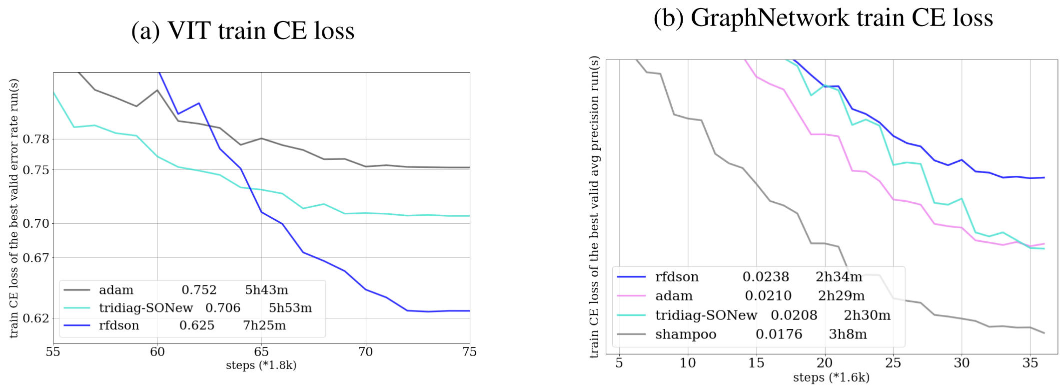


Figure 5: Train loss corresponding to the best validation runs in Figure 1 (a) VIT benchmark (b) GraphNetwork benchmark. We report the numbers and the training time in the legend. We observe that tridiag match or perform better than adam.

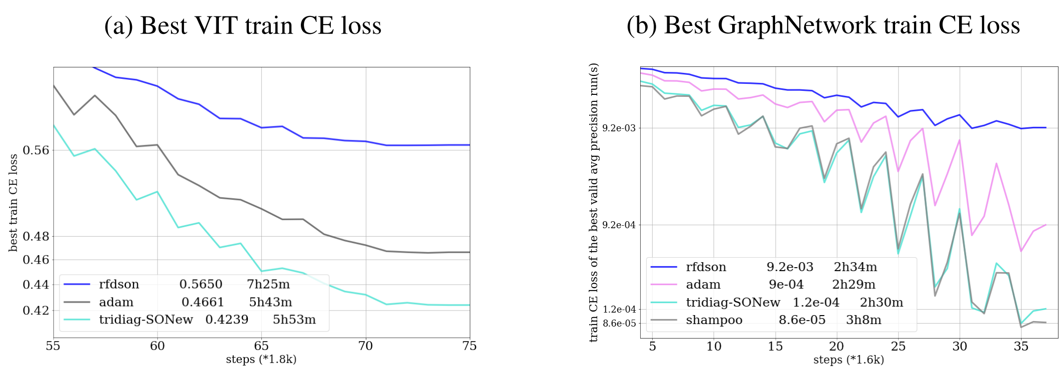


Figure 6: Best train loss achieved during hyperparam tuning. (a) VIT benchmark (b)GraphNetwork benchmark. We report the numbers and the training time in the legend. We observe that tridiag significantly outperforms adam, while being comparable to shampoo.

Table 8: **Comparison of rfdSON and tridiag-SONew in convex setting on three datasets. We optimize least square loss $\sum_t (y_t - w^T x_t)^2$ where w is the learnable parameter and (x_t, y_t) is the t^{th} training point. Reported numbers is the accuracy on the test set.**

Table 9: (a) Dataset stats

Dataset	# total points	dimension
a9a	32,561	123
gisette	6000	5000
mnist	11791	780

Table 10: (b) RFD-SON vs tridiag-SONew

Dataset	RFD-SON, m=2	RFD-SON, m=5	tridiag-SONew
a9a	83.3	83.6	84.6
gisette	96.1	96.2	96.6
mnist	93.2	94.5	96.5

Table 11: **Optimal hyperparams for Autoencoder Benchmark**

Table 12: (a) float32 experiments optimal hyperparamters

Baseline	β_1	β_2	ϵ	**lr**
SGD	0.99	0.91	8.37e-9	1.17e-2
Nesterov	0.914	0.90	3.88e-10	5.74e-3
Adagrad	0.95	0.90	9.96e-7	1.82e-2
Momentum	0.9	0.99	1e-5	6.89e-3
RMSProp	0.9	0.9	1e-10	4.61e-4
Adam	0.9	0.94	1.65e-6	3.75e-3
Diag-SONew	0.88	0.95	4.63e-6	1.18e-3
Shampoo	0.9	0.95	9.6e-9	3.70e-3
tridiag	0.9	0.96	1.3e-6	8.60e-3
band-4	0.88	0.95	1.5e-3	5.53e-3

Table 13: (b) bfloat16 experiments optimal hyperparamters

Baseline	β_1	β_2	ϵ	**lr**
SGD	0.96	0.98	2.80e-2	1.35e-2
Nesterov	0.914	0.945	8.48e-9	6.19e-3
Adagrad	0.95	0.93	2.44e-5	2.53e-2
Momentum	0.9	0.99	0.1	7.77e-3
RMSProp	0.9	0.9	2.53e-10	4.83e-4
Adam	0.9	0.94	3.03e-10	3.45e-3
Diag-SONew	0.9	0.95	4.07e-6	8.50e-3
Shampoo	0.85	0.806	6.58e-4	5.03e-3
ztridiag	0.83	0.954	1.78e-6	7.83e-3
band-4	0.9	0.96	1.52e-6	4.53e-3