Supplementary Information for Scalable Adaptive Stochastic Optimization Using Random Projections

A Computational Complexity

Table A.1: Comparison of computational complexity in big O notation between ADA-FULL, ADA-LR and RADAGRAD.

Operation	Line	ADA-FULL	ADA-LR	RADAGRAD
$\mathbf{\Pi g}_t$	3			$p \log \tau$
$\mathbf{G}_t = \mathbf{g}_t \mathbf{g}_t^\top$		p^2	p^2	$p\tau$
$\mathbf{G}_t \mathbf{\Pi}$	4		$p^2 \log \tau$	
QR-decomp	5		$\tau^2 p$	$\tau^2 p$
$\mathbf{Q}^\top \mathbf{G}_t$	6		τp^2	$\tau^2 p$
SVD	7	p^3	$\tau^2 p$	τ^3
$\mathbf{QW}$	8			$\tau^2 p$
$\boldsymbol{\beta}_{t+1} =$	10	p^2	τp	τp
Total		p^3	τp^2	$\tau^2 p$

B RADA-VR: RADAGRAD with variance reduction.

Algorithm 3 RADA-VR

Input: $\eta > 0$, $\delta \geq 0$, τ, S number of epochs, m iterations per epoch, initial $\boldsymbol{\beta}_0^1$

for $s = 1 \dots S$ **do**
  $\mu = \nabla \sum_{i=1}^n f_i(\boldsymbol{\beta}_0^s)$
  for $t = 1 \dots m-1$ **do**
    Compute VR gradient: $\mathbf{g}_t = \nabla f_t(\boldsymbol{\beta}_t^s) - \nabla f_t(\boldsymbol{\beta}_0^s) + \mu$
    Project: $\tilde{\mathbf{g}}_t = \mathbf{\Pi} \mathbf{g}_t$
    $\tilde{\mathbf{G}}_t = \tilde{\mathbf{G}}_{t-1} + \mathbf{g}_t \tilde{\mathbf{g}}_t^\top$
    $\mathbf{Q}_t, \mathbf{R}_t \leftarrow \texttt{qr\_update}(\mathbf{Q}_{t-1}, \mathbf{R}_{t-1}, \mathbf{g}_t, \tilde{\mathbf{g}}_t)$
    $\mathbf{B} = \tilde{\mathbf{G}}_t^\top \mathbf{Q}$
    $\mathbf{U}, \mathbf{\Sigma}, \mathbf{W} = \mathbf{B}$ {SVD}
    $\mathbf{V} = \mathbf{W} \tilde{\mathbf{Q}}^\top$
    $\boldsymbol{\beta}_{t+1}^s = \boldsymbol{\beta}_t^s - \eta \mathbf{V} (\mathbf{\Sigma}^{1/2} + \delta \mathbf{I})^{-1} \mathbf{V}^\top \mathbf{g}_t - \boldsymbol{\gamma}_t$
  end for
  $\boldsymbol{\beta}_0^{s+1} = \boldsymbol{\beta}_{t+1}^s$
end for

Output: $\boldsymbol{\beta}_m^S$

C Analysis

C.1 Regret bound for ADA-LR

The following proof is based on the proof for Theorem 7 in [9]. The key difference is that instead of having the square root and (pseudo-)inverse of the full matrix $\mathbf{G}_t$: $\mathbf{G}_t^{1/2}$ and $\mathbf{S}_t^\dagger$ we have the approximate square root and inverse based on the randomized SVD [15]): $\tilde{\mathbf{S}}_t = (\mathbf{Q}\mathbf{Q}^\top \mathbf{G}_t)^{1/2}$ and $\tilde{\mathbf{S}}_t^\dagger = (\mathbf{Q}\mathbf{Q}^\top \mathbf{G}_t)^{-1/2}$. Essentially we use the proximal function $\psi_t = \langle \mathbf{x}, \tilde{\mathbf{S}}_t \mathbf{x} \rangle$ or $\psi_t = \langle \mathbf{x}, \tilde{\mathbf{H}}_t \mathbf{x} \rangle$ where we set $\tilde{\mathbf{H}}_t = \delta \mathbf{I} + \tilde{\mathbf{S}}_t$. Here $\mathbf{Q}$ is the approximate basis for the range of the matrix $\mathbf{G}_t$ [15].

We first state the following facts about the relationship between $\mathbf{G}$ and $\tilde{\mathbf{G}}^{-1/2}$.

Lemma 3. *Defining* $\tilde{\mathbf{G}}^{-1/2} = (\mathbf{Q}\mathbf{Q}^\top\mathbf{G})^{-1/2}$ *we have*

(I) $\tilde{\mathbf{G}}^{-1/2}\mathbf{G} = (\mathbf{G}^{-1}(\mathbf{Q}\mathbf{Q}^\top)\mathbf{G}^2)^{1/2}$,

(II) $\mathrm{tr}((\mathbf{G}^{-1}(\mathbf{Q}\mathbf{Q}^\top)\mathbf{G}^2)^{1/2}) = \mathrm{tr}(\tilde{\mathbf{G}}^{1/2})$.

We also require the following Lemma which bounds the sequence of proximal terms by the trace of the final $\tilde{\mathbf{G}}^{-1/2}$.

Lemma 4 (Based on Lemma 10 in [9])**.**

$$\sum_{t=1}^{T}\langle \mathbf{g}_t, \tilde{\mathbf{G}}_t^{-1/2}\mathbf{g}_t\rangle \leq 2\sum_{t=1}^{T}\langle \mathbf{g}_t, \tilde{\mathbf{G}}_T^{-1/2}\mathbf{g}_t\rangle = 2\mathrm{tr}(\tilde{\mathbf{G}}_T^{1/2}). \tag{3}$$

We are now ready to prove Proposition 2.

Proof of Proposition 2. Inspecting Lemma 6:

$$R(T) \leq \frac{1}{\eta}\psi_T(\boldsymbol{\beta}^{\text{opt}}) + \frac{\eta}{2}\sum_{t=1}^{T}\|f_t'(\boldsymbol{\beta}_t)\|^2_{\psi^*_{T-1}},$$

we first bound the term $\sum_{t=1}^{T}\|f_t'(\boldsymbol{\beta}_t)\|^2_{\psi^*_{T-1}}$.

From [9, Proof of Theorem 7] we have that the squared dual norm associated with ψ_t is

$$\|\mathbf{x}\|^2_{\psi_t^*} = \langle \mathbf{x}, (\delta\mathbf{I} + (\mathbf{Q}\mathbf{Q}^\top\mathbf{G}_t)^{1/2})^{-1}\mathbf{x}\rangle$$

and thus it is clear that $\|\mathbf{g}_t\|^2_{\psi_t^*} \leq \langle \mathbf{g}_t, (\mathbf{Q}\mathbf{Q}^\top\mathbf{G}_t)^{-1/2}\mathbf{g}_t\rangle$. Lemma 8 shows that $\|\mathbf{g}_t\|^2_{\psi^*_{t-1}} \leq \langle \mathbf{g}_t, \tilde{\mathbf{S}}_t\mathbf{g}_t\rangle$ as long as $\delta \geq \|\mathbf{g}_t\|_2$. Lemma 4 then implies that

$$\sum_{t=1}^{T}\|f_t'(\boldsymbol{\beta}_t)\|^2_{\psi^*_{T-1}} \leq 2\mathrm{tr}(\tilde{\mathbf{G}}_T^{1/2}).$$

We now bound $2\mathrm{tr}(\tilde{\mathbf{G}}_T^{1/2})$ by $2(\mathrm{tr}(\mathbf{G}_T^{1/2}) + \tau\sqrt{\epsilon})$:

$$\begin{aligned}
\mathrm{tr}(\tilde{\mathbf{G}}_T^{1/2}) - \mathrm{tr}(\mathbf{G}_T^{1/2}) &= \mathrm{tr}(\tilde{\mathbf{G}}_T^{1/2} - \mathbf{G}_T^{1/2}) && (4)\\
&= \sum_{j=1}^{\tau}\left(\lambda_j(\tilde{\mathbf{G}}_T^{1/2}) - \lambda_j(\mathbf{G}_T^{1/2})\right) - \sum_{j=\tau+1}^{p}\lambda_j(\mathbf{G}_T^{1/2}) && (5)\\
&\leq \sum_{j=1}^{\tau}\left(\lambda_1(\tilde{\mathbf{G}}_T^{1/2}) - \lambda_1(\mathbf{G}_T^{1/2})\right) && (6)
\end{aligned}$$

since $\lambda_j(\tilde{\mathbf{G}}_T) = 0,\ \forall j > \tau$.

Now, using the reverse triangle inequality and Theorem 5 we obtain

$$\begin{aligned}
\sum_{j=1}^{\tau}\left(\lambda_1(\tilde{\mathbf{G}}_T^{1/2}) - \lambda_1(\mathbf{G}_T^{1/2})\right) &\leq \sum_{j=1}^{\tau}\|\tilde{\mathbf{G}}_T^{1/2} - \mathbf{G}_T^{1/2}\|_2 && (7)\\
&\leq \sum_{j=1}^{\tau}\sqrt{\epsilon} && (8)\\
&\leq \tau\sqrt{\epsilon}. && (9)
\end{aligned}$$

It remains to show that $\psi_T(\boldsymbol{\beta}^{\text{opt}})$ in Lemma 6 is bounded by $\left(\delta + \sqrt{\epsilon} + \text{tr}(\mathbf{G}_T^{1/2})\right) \|\boldsymbol{\beta}^{\text{opt}}\|^2$ to get the statement of Theorem 2:

$$\begin{aligned}
\psi_T(\boldsymbol{\beta}^{\text{opt}}) &= \langle \boldsymbol{\beta}^{\text{opt}}, \delta\mathbf{I} + (\mathbf{Q}\mathbf{Q}^\top \mathbf{G}_T)^{1/2} \boldsymbol{\beta}^{\text{opt}} \rangle \\
&\leq \|\boldsymbol{\beta}^{\text{opt}}\|^2 \|(\mathbf{Q}\mathbf{Q}^\top \mathbf{G}_T)^{1/2}\|_2 + \delta \|\boldsymbol{\beta}^{\text{opt}}\|^2 \\
&\leq \|\boldsymbol{\beta}^{\text{opt}}\|^2 \left(\sqrt{\epsilon} + \|\mathbf{G}_T^{1/2}\|\right) + \delta \|\boldsymbol{\beta}^{\text{opt}}\|^2 \\
&\leq \|\boldsymbol{\beta}^{\text{opt}}\|^2 \left(\sqrt{\epsilon} + \text{tr}(\mathbf{G}_T^{1/2})\right) + \delta \|\boldsymbol{\beta}^{\text{opt}}\|^2
\end{aligned}$$

where we again use the reverse triangle inequality and Theorem 5 as above.

Finally, plugging this into the statement of Lemma 6 and setting $\eta = \|\boldsymbol{\beta}^{\text{opt}}\|_2$ (as in Corollary 11 in [9]) we get the expression for the regret of ADA-LR as stated in Theorem 2. □

C.2 Proofs of supporting results

Proof of Lemma 3. By direct computation we have for (I)

$$\begin{aligned}
\tilde{\mathbf{G}}^{-1/2}\mathbf{G} &= (\mathbf{Q}\mathbf{Q}^\top \mathbf{G})^{-1/2}\mathbf{G} \\
&= ((\mathbf{Q}\mathbf{Q}^\top \mathbf{G})^{-1}\mathbf{G}^2)^{1/2} \\
&= (\mathbf{G}^{-1}(\mathbf{Q}\mathbf{Q}^\top)^{-1}\mathbf{G}^2)^{1/2} \\
&= (\mathbf{G}^{-1}(\mathbf{Q}\mathbf{Q}^\top)\mathbf{G}^2)^{1/2}.
\end{aligned}$$

and for (II)

$$\begin{aligned}
\text{tr}((\mathbf{G}^{-1}(\mathbf{Q}\mathbf{Q}^\top)\mathbf{G}^2)^{1/2}) &= \text{tr}((\mathbf{Q}^\top \mathbf{G}\mathbf{Q})^{1/2}) \\
&= \text{tr}((\mathbf{Q}\mathbf{Q}^\top \mathbf{G})^{1/2}) \\
&= \text{tr}(\tilde{\mathbf{G}}^{1/2}).
\end{aligned}$$

□

Proof of Lemma 4. We set up the following proof by induction. In the base case:

$$\langle \mathbf{g}_1, \tilde{\mathbf{G}}_1^{-1/2}, \mathbf{g}_1 \rangle = \text{tr}(\tilde{\mathbf{G}}_1^{-1/2} \mathbf{g}_1 \mathbf{g}_1^\top) = \text{tr}(\tilde{\mathbf{G}}_1^{1/2}) \leq 2\text{tr}(\tilde{\mathbf{G}}_1^{1/2}),$$

where we have used (II).

Now, assuming that the lemma is true for $T-1$, we get:

$$\sum_{t=1}^{T} \langle g_t, \tilde{\mathbf{G}}_t^{-1/2}, g_t \rangle \leq 2 \sum_{t=1}^{T} \langle \mathbf{g}_t, \tilde{\mathbf{G}}_{T-1}^{-1/2} \mathbf{g}_t \rangle + \langle \mathbf{g}_T, \tilde{\mathbf{G}}_T^{-1/2} \mathbf{g}_T \rangle.$$

Now using that $\tilde{\mathbf{G}}_{T-1}^{-1/2}$ does not depend on t and (II):

$$\sum_{t=1}^{T-1} \langle g_t, \tilde{\mathbf{G}}_{T-1}^{-1/2} g_t \rangle = \text{tr}(\tilde{\mathbf{G}}_{T-1}^{-1/2} \mathbf{G}_{T-1}) = \text{tr}(\tilde{\mathbf{G}}_{T-1}^{1/2}).$$

Therefore we get

$$\sum_{t=1}^{T} \langle g_t, \tilde{\mathbf{G}}_t^{-1/2} g_t \rangle \leq 2\text{tr}(\tilde{\mathbf{G}}_{T-1}^{1/2}) + \langle \mathbf{g}_T, \tilde{\mathbf{G}}_T^{-1/2} \mathbf{g}_T \rangle. \tag{10}$$

We can rewrite

$$\text{tr}(\tilde{\mathbf{G}}_{T-1}^{1/2}) = \text{tr}(\left(\mathbf{Q}_{T-1}\mathbf{Q}_{T-1}^\top \mathbf{G}_T - \mathbf{Q}_{T-1}\mathbf{Q}_{T-1}^\top \mathbf{g}_T \mathbf{g}_T^\top\right)^{1/2}) \tag{11}$$

Now since $\text{range}(\mathbf{Q}_{T-1}) \subset \text{range}(\mathbf{Q}_T)$ and Proposition 8.5 in [15] we can use Lemma 7 with $\nu = 1$ and $\mathbf{g} = \mathbf{g}_t$ to obtain:

$$2\text{tr}(\tilde{\mathbf{G}}_{T-1}^{1/2}) + \langle \mathbf{g}_T, \tilde{\mathbf{G}}_T^{-1/2}, \mathbf{g}_T \rangle \leq 2\text{tr}(\tilde{\mathbf{G}}_T^{1/2}) \tag{12}$$

□

D Supporting Results

Theorem 5 (SRFT approximation error (Theorem 11.2 in [15])). *Defining* $\epsilon = \sqrt{1 + 7p/\tau} \cdot \sigma_{k+1}$ *the following holds with failure probability at most* $O\left(k^{-1}\right)$

$$\left\|\mathbf{G}_t - \mathbf{Q}\mathbf{Q}^\top \mathbf{G}_t\right\|_2 \leq \epsilon, \tag{13}$$

where σ_{k+1} *is the* k*th largest singular value of* $\mathbf{G}_t$*, and* $4\left[\sqrt{k} + \sqrt{8\log(kn)}\right]^2 \leq \tau \leq p$*.*

Lemma 6 (Proposition 2 from [9])**.**

$$R(T) := \sum_{t=1}^{T} f_t(\boldsymbol{\beta}_t) + \varphi(\boldsymbol{\beta}_t) - f_t(\boldsymbol{\beta}^{opt}) - \varphi(\boldsymbol{\beta}^{opt}) \leq \frac{1}{\eta}\psi_T(\boldsymbol{\beta}^{opt}) + \frac{\eta}{2}\sum_{t=1}^{T} \|f_t'(\boldsymbol{\beta}_t)\|^2_{\psi^*_{T-1}}$$

Lemma 7 (Lemma 8 from [9])**.** *Let* $\mathbf{B} \succeq 0$*. For any* ν *such that* $\mathbf{B} - \nu\mathbf{g}\mathbf{g}^\top \succeq 0$ *the following holds*

$$2\mathrm{tr}((\mathbf{B} - \nu\mathbf{g}\mathbf{g}^\top)^{1/2}) \leq 2\mathrm{tr}(\mathbf{B}^{1/2}) - \nu\mathrm{tr}(\mathbf{B}^{1/2}\mathbf{g}\mathbf{g}^\top)$$

Lemma 8 (Lemma 9 from [9])**.** *Let* $\delta \geq \|\mathbf{g}\|_2$ *and* $\mathbf{A} \succeq 0$*, then*

$$\langle \mathbf{g}, (\delta\mathbf{I} + \mathbf{A}^{1/2})^{-1}\mathbf{g}\rangle \leq \langle \mathbf{g}, ((\mathbf{A} + \mathbf{g}\mathbf{g}^\top)^\dagger)^{1/2}\mathbf{g}\rangle$$

References

[1] Z. Allen-Zhu and E. Hazan. Variance reduction for faster non-convex optimization. In *Proceedings of the 33rd International Conference on Machine Learning*, 2016.

[2] S.-I. Amari. Natural gradient works efficiently in learning. *Neural computation*, 10(2):251–276, 1998.

[3] D. Balduzzi. Deep online convex optimization with gated games. *arXiv preprint arXiv:1604.01952*, 2016.

[4] D. Balduzzi and M. Ghifary. Strongly-typed recurrent neural networks. In *Proceedings of the 33rd International Conference on Machine Learning*, 2016.

[5] R. H. Byrd, S. Hansen, J. Nocedal, and Y. Singer. A stochastic quasi-newton method for large-scale optimization. *arXiv preprint arXiv:1401.7020*, 2014.

[6] Y. N. Dauphin, H. de Vries, J. Chung, and Y. Bengio. Rmsprop and equilibrated adaptive learning rates for non-convex optimization. *arXiv preprint arXiv:1502.04390*, 2015.

[7] A. Defazio, F. Bach, and S. Lacoste-Julien. Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in Neural Information Processing Systems*, 2014.

[8] G. Desjardins, K. Simonyan, R. Pascanu, et al. Natural neural networks. In *Advances in Neural Information Processing Systems*, pages 2062–2070, 2015.

[9] J. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. *The Journal of Machine Learning Research*, 12:2121–2159, 2011.

[10] J. C. Duchi, M. I. Jordan, and H. B. McMahan. Estimation, optimization, and parallelism when data is sparse. In *Advances in Neural Information Processing Systems*, 2013.

[11] G. H. Golub and C. F. Van Loan. *Matrix computations*, volume 3. JHU Press, 2012.

[12] A. Gonen and S. Shalev-Shwartz. Faster sgd using sketched conditioning. *arXiv preprint arXiv:1506.02649*, 2015.

[13] Y. Gong, S. Kumar, H. Rowley, and S. Lazebnik. Learning binary codes for high-dimensional data using bilinear projections. In *Proceedings of CVPR*, pages 484–491, 2013.

[14] R. Grosse and R. Salakhudinov. Scaling up natural gradient by sparsely factorizing the inverse fisher matrix. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 2304–2313, 2015.

[15] N. Halko, P. G. Martinsson, and J. A. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review*, 53(2):217–288, 2011.

[16] C. Heinze, B. McWilliams, and N. Meinshausen. Dual-loco: Distributing statistical estimation using random projections. In *Proceedings of AISTATS*, 2016.

[17] C. Heinze, B. McWilliams, N. Meinshausen, and G. Krummenacher. Loco: Distributing ridge regression with random projections. *arXiv preprint arXiv:1406.3469*, 2014.

[18] T. Hofmann, A. Lucchi, S. Lacoste-Julien, and B. McWilliams. Variance reduced stochastic gradient descent with neighbors. In *Advances in Neural Information Processing Systems*, 2015.

[19] R. Johnson and T. Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems*, pages 315–323, 2013.

[20] N. S. Keskar and A. S. Berahas. adaQN: An Adaptive Quasi-Newton Algorithm for Training RNNs. Nov. 2015.

[21] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[22] A. Lucchi, B. McWilliams, and T. Hofmann. A variance reduced stochastic newton method. *arXiv preprint arXiv:1503.08316*, 2015.

[23] H. Luo, A. Agarwal, N. Cesa-Bianchi, and J. Langford. Efficient second order online learning via sketching. *arXiv preprint arXiv:1602.02202*, 2016.

[24] M. W. Mahoney. Randomized algorithms for matrices and data. Apr. 2011. arXiv:1104.5557v3 [cs.DS].

[25] M. P. Marcus, M. A. Marcinkiewicz, and B. Santorini. Building a large annotated corpus of english: The penn treebank. *Computational linguistics*, 19(2):313–330, 1993.

[26] J. Martens and R. Grosse. Optimizing neural networks with kronecker-factored approximate curvature. In *Proceedings of the 32nd International Conference on Machine Learning*, 2015.

[27] B. McWilliams, G. Krummenacher, M. Lucic, and J. M. Buhmann. Fast and robust least squares estimation in corrupted linear models. In *Advances in Neural Information Processing Systems*, volume 27, 2014.

[28] B. Neyshabur, R. R. Salakhutdinov, and N. Srebro. Path-sgd: Path-normalized optimization in deep neural networks. In *Advances in Neural Information Processing Systems*, pages 2413–2421, 2015.

[29] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *EMNLP*, volume 14, pages 1532–1543, 2014.

[30] L. Zhang, M. Mahdavi, R. Jin, T. Yang, and S. Zhu. Recovering optimal solution by dual random projection. *arXiv preprint arXiv:1211.3046*, 2012.