A. Appendix

A.1. Proof of Theorem 2.1

Proof. We prove this theorem by way of the following two lemmas:

Lemma A.1. *Let $i \in \{1, 2, ..., L\}$. Then for all $j \in \{1, 2, ..., L\}$ s.t. $i \notin A_j$ we have*

$$\mathbb{E}\left[v_i S_{y_j}(V)^\top\right] = F_i J_{y_j}^{x_i}$$

Lemma A.2. *Let $j \in \{1, 2, ..., L\}$. Then for all $i \in \{1, 2, ..., L\}$ s.t. $i \notin A_j$ we have*

$$\mathbb{E}[S_{y_i}(V) S_{y_j}(V)^\top] = H_{y_i,y_j}^f$$

Taking $i = j = 1$ in Lemma A.2 gives:

$$\mathbb{E}[S(V)S(V)^\top] = \mathbb{E}\left[(S_{y_k}(V) S_{y'_k}(V))^\top\right] = H_{y_1,y_1}^f \equiv H$$

□

Proof of Lemma A.1. One method of proof is to use structural induction on the computation graph. Instead, we will assume that if $j \in A_i$ then $j < i$ (which can be done without loss of generality since it it's always possible to reindex the nodes of the graph in this way), and proceed by standard induction on j, starting at $j = L$ and proceeding backwards towards $j = 1$.

The base case occurs when $j = L$. Letting i be s.t. $i \notin A_j$ it must be the case that $i = L$. So we have

$$\mathbb{E}\left[v_L S_{y_L}(V)^\top\right] = \mathbb{E}[v_L 0] = 0 = F_L 0 = F_L J_{y_L}^{x_L}$$

where we used the fact that $J_{y_i}^{x_i} = 0$.

For the inductive case suppose that $j \in \{1, ..., L-1\}$ and that the claim holds for strictly larger j's. Then,

$$\begin{aligned}
\mathbb{E}\left[v_i S_{y_j}(V)^\top\right] &= \mathbb{E}\left[v_i \left(\sum_{k\in C_j} R_{k,j}^\top S_{x_k}(V)\right)^\top\right] \\
&= \sum_{k\in C_j} \mathbb{E}\left[v_i S_{x_k}(V)^\top\right] R_{k,j} = \sum_{k\in C_j} F_i J_{x_k}^{x_i} R_{k,j} \\
&= F_i \sum_{k\in C_j} J_{x_k}^{x_i} R_{k,j} = F_i J_{y_j}^{x_i}
\end{aligned}$$

where the last line follows from eqn. 10, and second line follows from the fact that $\mathbb{E}[v_i S_{x_k}(V)^\top] = F_i J_{x_i,x_k}$ which can be proven as follows:

$$\begin{aligned}
\mathbb{E}[v_i S_{x_k}(V)^\top] &= \mathbb{E}\left[v_i \left(F_k^\top v_k + J_{x_k}^{y_k\,\top} S_{y_k}(V)\right)^\top\right] \\
&= \mathbb{E}\left[v_i v_k^\top\right] F_j + \mathbb{E}[v_i S_{y_k}(V)] J_{x_k}^{y_k} \\
&= (\delta_{ik} I) F_k + F_i J_{y_k}^{x_i} J_{x_k}^{y_k} \\
&= F_i(\delta_{ik} I + (1 - \delta_{ik}) J_{x_k}^{x_i}) = F_i J_{x_k}^{x_i}
\end{aligned}$$

where the third line follows from the inductive hypothesis (which applies since $k \in C_j \Rightarrow k > j$) and we have used the identity $J_{x_i}^{x_i} = I$, and $J_{y_k}^{x_i} = 0$ when $i = k$, and otherwise $J_{y_k}^{x_i} J_{x_k}^{y_k} = J_{x_k}^{x_i}$ when $i \neq k$. □

Proof of Lemma A.2. As in the previous lemma we proceed by induction on j, going from $j = L$ down to $j = 1$.

The base case is trivial since even without taking expectations we have $S_{y_L}(V) S_{y_j}(V)^\top = \mathrm{vec}(0) S_{y_j}(V)^\top = 0 = H_{y_L,y_L}^{y_L} = H_{y_L,y_L}^f$ (since $f = y_L$).

For the inductive case suppose that $j \in \{1, ..., L-1\}$ and that the claim holds for strictly larger j's. Then noting that if i is k's input (i.e., $k \in C_i$), then $k > i$ and $k \notin A_j$ (which uses the additional facts that $i \notin A_j$ and the computational graph contains no dependency cycles), we have

$$\begin{aligned}
\mathbb{E}\left[S_{y_i}(V) S_{y_j}(V)^\top\right] &= \mathbb{E}\left[\sum_{k\in C_i} R_{k,i}^\top S_{x_k}(V) S_{y_j}(V)^\top\right] \\
&= \sum_{k\in C_i} R_{k,i}^\top \,\mathbb{E}\left[S_{x_k}(V) S_{y_j}(V)^\top\right] \\
&= \sum_{k\in C_i} R_{k,i}^\top H_{x_k,y_j}^f = H_{y_i,y_j}^f
\end{aligned}$$

where the third line follows from eqn. 5, and the second line follows from the fact that $\mathbb{E}\left[S_{x_k}(V) S_{y_j}(V)^\top\right] = H_{x_k,y_j}^f$ which can be proven as follows:

$$\begin{aligned}
\mathbb{E}\left[S_{x_k}(V) S_{y_j}(V)^\top\right] &= \mathbb{E}\left[(F_k^\top v_k + J_{x_k}^{y_k\,\top} S_{y_k}(V)) S_{y_j}(V)^\top\right] \\
&= F_k^\top \,\mathbb{E}\left[v_k S_{y_j}(V)^\top\right] + J_{x_k}^{y_k\,\top} \,\mathbb{E}\left[S_{y_k}(V) S_{y_j}(V)^\top\right] \\
&= F_k^\top F_k J_{y_j}^{x_k} + J_{x_k}^{y_k\,\top} H_{y_k,y_j}^f \\
&= M_k J_{y_j}^{x_k} + J_{x_i}^{y_k\,\top} H_{y_k,y_j}^f = H_{x_k,y_j}^f
\end{aligned}$$

where the third line follows from Lemma A.1 and the inductive hypothesis (which applies since $k \in C_j \Rightarrow k > j$), and the forth from eqn. 6. □

Proof of Theorem 2.2. The proof proceeds along very similar lines to Theorem 2.1 and is thus omitted. □

A.2. Proof of Theorem 4.1 (variance inequality)

Proof. In the general case we have:

$$\mathrm{Var}_G\left[H_{ii}^{A,B}\right] = (A_i^\top A_i)(B_i^\top B_i) + H_{ii}^2$$

and so in the more specific case that $A = B = \tilde{S}$ we have:

$$\mathrm{Var}_G\left[H_{ii}^{\tilde{S},\tilde{S}}\right] = (\tilde{S}_i^\top \tilde{S}_i)^2 + H_{ii}^2 = 2H_{ii}^2$$

But if we apply the the Cauchy-Swartz inequality we have:

$$\begin{aligned}
\mathrm{Var}_G\left[H_{ii}^{A,B}\right] &= (A_i^\top A_i)(B_i^\top B_i) + H_{ii}^2 \\
&\geq \left(A_i^\top B_i\right)^2 + H_{ii}^2 \\
&= 2H_{ii}^2 = \mathrm{Var}_G\left[H_{ii}^{\tilde{S},\tilde{S}}\right]
\end{aligned}$$

□

A.3. Proof of Theorem 6.1 and Lemma 6.2 (circuit complexity results)

Proof of Theorem 6.1. Suppose by contradiction that there is a bounded depth arithmetic circuit family that computes the diagonal of $f(y) = 1/2y^\top W^\top ZWy$ with $O(n^2)$ edges. If follows trivially from Lemma 6.2 there must also exist a circuit family of edge count $O(n^2)$ which computes the product of 2 $n \times n$ input matrices, which contradicts a result of Raz and Shpilka (2001) which says that such a circuit family must have an edge count which is superlinear in n^2. □

Proof of Lemma 6.2. This result is similar to one proved by Raz and Shpilka (2001) which concerned the computation the trace of the product of 3 arbitrary matrices. We will use adopt their proof technique here.

Construct $W \equiv [P^\top Q]^\top$ from the input matrices P and Q (which can be done with $2n^2$ edges).

By hypothesis there exists an arithmetic circuit with arbitrary fan-in gates, which given P, Q and Z as input, will compute the diagonal of the Hessian of f, which is $W^\top ZW$. Append to this circuit a single sum gate which computes the sum the outputs, thus obtaining the trace of $W^\top ZW$ and adding a single layer of depth and n edges. Then, using a result of Walter and Strassen (1983), there is also an arithmetic circuit for computing all the derivatives of the function computed by this circuit (i.e. trace$(W^\top ZW)$) w.r.t. Z which has twice the depth and three times the size of the original circuit (the derivative circuit works by performing what amounts to automatic-differentiation).

But note that:

$$\frac{d\,\text{trace}(W^\top ZW)}{dZ} = \frac{d\,\text{trace}((WW^\top)^\top Z)}{dZ}$$
$$= WW^\top = \begin{bmatrix} PP^\top & PQ \\ Q^\top P^\top & Q^\top Q \end{bmatrix}$$

where we have used the well-known facts that $\frac{d\,\text{trace}(AB)}{dB} = A^\top$ and that trace is invariant under cyclic permutations of matrix products.

By taking the upper-right corner of this output matrix and discarding the rest, the circuit thus computes the product PQ. □

A.4. On the Hessian estimates used in Rifai et al. (2011)

In Rifai et al. (2011) the authors estimate the Frobenius norm of the Hessian via the 0 variance limit of a stochastic finite-difference formula:

$$\|H\|_F^2 = \lim_{\sigma\to 0} \frac{1}{\sigma^2} \mathbb{E}_w \left[\|\nabla f(y_1 + \sigma w) - f(y_1)\|^2\right]$$

where $w \sim \text{Normal}(0, I)$.

A simpler derivation of this result than that which appears in Rifai et al. (2011) is:

$$\lim_{\sigma\to 0} \frac{1}{\sigma^2} \mathbb{E}_w \left[\|\nabla f(y_1 + \sigma w) - f(y_1)\|^2\right]$$
$$= \mathbb{E}_w \left[\left\|\lim_{\sigma\to 0} \frac{\nabla f(y_1 + \sigma w) - f(y_1)}{\sigma}\right\|^2\right]$$
$$= \mathbb{E}_w \left[\|Hw\|^2\right] = \mathbb{E}_w[(Hw)^\top Hw]$$
$$= \mathbb{E}_w[w^\top HHw] = \mathbb{E}_w[\text{trace}(Hww^\top H)]$$
$$= \text{trace}(H\,\mathbb{E}_w[ww^\top]H) = \text{trace}(HIH) = \|H\|_F^2$$

where we have used the well-known identity for Hessian-vector products: $\lim_{\sigma\to 0} \frac{\nabla f(y_1+\sigma w) - f(y_1)}{\sigma} = Hw$ and the property that $\mathbb{E}_w[ww^\top] = I$. Moreover, this derivation suggests how one can forgo the unreliable finite differences approximation in favor of Hessian-vector products computed efficiently and exactly via automatic differentiation-type methods (e.g. Pearlmutter, 1994). That is, we can sample w (from any distribution satisfying $\mathbb{E}_w[ww^\top] = I$, we are not restricted to use Normal$(0, I)$), compute $z = Hw$, and then obtain our unbiased estimate of $\|H\|_F^2$ as $z^\top z$.

Note that the estimator $\|\hat{H}\|_F^2$, where $\hat{H}$ is some unbiased estimator of H (e.g. obtained from CP), won't be unbiased in general. However, an unbiased estimator can be obtained using the techniques of CP by sampling an appropriate V, computing $z = S(V) = \tilde{S}v$ (using the notation of section 2.6), and then taking $z^\top Hz$. That this is unbiased can be easily checked:

$$\mathbb{E}_v\left[(\tilde{S}v)^\top H(\tilde{S}v)\right] = \mathbb{E}_v\left[\text{trace}((\tilde{S}v)^\top H(\tilde{S}v))\right]$$
$$= \mathbb{E}_v\left[\text{trace}(v^\top \tilde{S}^\top H\tilde{S}v)\right] = \text{trace}(H\tilde{S}\,\mathbb{E}_v\left[vv^\top\right]\tilde{S}^\top)$$
$$= \text{trace}(H\tilde{S}I\tilde{S}^\top) = \text{trace}(HH) = \|H\|_F^2$$