A Appendix

We now provide additional details for our results in Section 2 of the paper.

Lemma 1. *Let A, B be full column rank matrices of size $n \times d$, with $\log(n) = d^{o(1)}$. Let S be an SRHT with*[1] $m = \tilde{O}((d + \log(1/\delta))/\epsilon^2)$ *rows. For any matrix B of size $n \times d$ we have*

$$\|X\|_2 = \|(SA)^\dagger SB\|_2 \lesssim \|A^\dagger B\|_2 + \epsilon\|\Sigma^{-1}\|_2(\sqrt{(1 + d/k)(\|B\|_2^2 + \|B\|_F^2/k)}),$$

with probability $1 - 1/poly(d)$.

Proof. From Equation 5 in the body and the triangle inequality, we have

$$(SA)^\dagger SB = V\Sigma^{-1}(\sum_{k=0}^{\infty} T^K)U^{\mathrm{T}}S^{\mathrm{T}}SB \tag{1}$$

$$\begin{aligned}
\|(SA)^\dagger SB\|_2 &\leq \|A^\dagger B\|_2 + \|(SA)^\dagger SB - A^\dagger B\|_2 && (2)\\
&\leq \|A^\dagger B\|_2 + \|V\Sigma^{-1}(\sum_{k=0}^{\infty} T^k)U^{\mathrm{T}}S^{\mathrm{T}}SB - V\Sigma^{-1}U^{\mathrm{T}}B\|_2 && (3)\\
&\leq \|A^\dagger B\|_2 + \|\Sigma^{-1}\|_2 \sum_{k=0}^{\infty} \epsilon^k \|U^{\mathrm{T}}S^{\mathrm{T}}SB - U^{\mathrm{T}}B\|_2 && (4)\\
&\leq \|A^\dagger B\|_2 + \|\Sigma^{-1}\|_2 \frac{\epsilon}{1-\epsilon}\sqrt{(1 + d/k)(\|B\|_2^2 + \|B\|_F^2/k)} && (5)
\end{aligned}$$

where we used equation 8 in the body in the last step. □

A.1 AD

Let us manually derive the AD for the least squares regression problem.

$$\begin{aligned}
\mathrm{LLS}(A, b) &= \mathrm{LS}(A^{\mathrm{T}}A, A^{\mathrm{T}}b)\\
&= \mathrm{LS}(M, m)\\
(M, m) &\equiv (A^{\mathrm{T}}A, A^{\mathrm{T}}b)\\
B &= A^{\mathrm{T}}\\
(B_1, B_2) &= (B, B)\\
C &= B_1 A\\
d &= B_2 b\\
\bar{A}_M &= A\bar{M} + A\bar{M}^{\mathrm{T}}\\
\bar{A}_m &= b\bar{m}^{\mathrm{T}}\\
\bar{A} &= \bar{A}_M + \bar{A}_m\\
&= A\bar{M} + A\bar{M}^{\mathrm{T}} + b\bar{m}^{\mathrm{T}}\\
\bar{b} &= A\bar{m}\\
(\bar{M}, \bar{m}) &= \mathcal{J}^{\mathrm{T}}(\mathrm{LS})(M, m)(\bar{y})\\
&= (-\bar{m}y^{\mathrm{T}}, \mathrm{LS}(M^{\mathrm{T}}, \bar{y})\\
&= (-\mathrm{LS}(A^{\mathrm{T}}A, \bar{y})y^{\mathrm{T}}, \mathrm{LS}(A^{\mathrm{T}}A, \bar{y}))
\end{aligned}$$

This gives us the final reverse mode AD:

$$\bar{A} = -A(A^{\mathrm{T}}A)^{-1}\bar{y}y^{\mathrm{T}}) - Ay\bar{y}^{\mathrm{T}}(A^{\mathrm{T}}A)^{-1} + by^{\mathrm{T}}(A^{\mathrm{T}}A)^{-1} \tag{6}$$

$$\bar{b} = A(A^{\mathrm{T}}A)^{-1}\bar{y} \tag{7}$$

[1]For a function f, we use the notation $\tilde{O}(f)$ to denote $f \cdot \mathrm{polylog}(f)$.

A.2 Approximation bounds

Let us derive some additional bounds which were missing in the main paper:

$$\|Ay\bar{y}^{\mathrm{T}}M^{-1} - Ay_D\bar{y}^TM_S^{-1}\|_F \leq \|Ay\bar{y}^{\mathrm{T}}M^{-1} - Ay_D\bar{y}^TM^{-1}\|_F + \|Ay_D\bar{y}^{\mathrm{T}}(M - M_S^{-1})\|_F \tag{8}$$

$$\leq \|U(I - (U^{\mathrm{T}}S^{\mathrm{T}}SU)^{-1})U^{\mathrm{T}}b\|_F + \epsilon\|Ay_D\|\|\bar{y}\|\|\Sigma^{-1}\|_2\|\Sigma^{-1}\|_F \tag{9}$$

$$\lesssim \epsilon(\|b\|_2 + \|b\|_2\|\bar{y}\|_2\|\Sigma^{-1}\|_2\|\Sigma^{-1}\|_F) \tag{10}$$

Table 1: Cheat sheet to derive AD.

Original	**Forward Transform**	**Reverse Transform**
$z = a + b$	$\dot{z} = \dot{a} + \dot{b}$	$(\bar{a}, \bar{b}) = (\bar{z}, \bar{z})$
$z = ab$	$\dot{z} = \dot{a}b + a\dot{b}$	$(\bar{a}, \bar{b}) = (\bar{z}b, a\bar{z})$
$(z_1, z_2) = (a, a)$	$(\dot{z}_1, \dot{z}_2) = (\dot{a}, \dot{a})$	$\bar{a} = \bar{z}_1 + \bar{z}_2$
$Y = AXB$	$\dot{Y} = A\dot{Y}B$	$\bar{X} = A^{\mathrm{T}}\bar{Y}B^{\mathrm{T}}$
$y = \mathrm{LS}(M, m)$	$\dot{y} = \mathcal{J}\,\mathrm{LS}(M, m)(y, \dot{M}, \dot{m})$ $= \mathrm{LS}(M, \dot{m} - \dot{M}y)$	$(\bar{M}, \bar{m}) = \mathcal{J}^{\mathrm{T}}\,\mathrm{LS}(M, m)(y, \bar{y})$ $= (-\bar{m}y^{\mathrm{T}}, \mathrm{LS}(M^{\mathrm{T}}, \bar{y}))$

Table 2: Forward mode AD Transformations.

Type	**Primal**	**Forward Transform**
Regular	$y = \mathrm{LLS}(A, b)$	$\dot{y} = \mathcal{J}\,\mathrm{LLS}(A, b)(y, \dot{A}, \dot{b})$ $= \mathrm{LS}(A^{\mathrm{T}}A, \dot{A}^{\mathrm{T}}b + A^{\mathrm{T}}\dot{b} - (\dot{A}^{\mathrm{T}}A + A^{\mathrm{T}}\dot{A})y)$
"Diff + Sketch"	$y_D = \mathrm{LLS}(A, b, S)$	$\dot{y}_D = \mathcal{J}\,\mathrm{LLS}(A, b)(y_D, \dot{A}, \dot{b}, S)$ $= \mathrm{LS}(A^{\mathrm{T}}S^{\mathrm{T}}SA, \dot{A}^{\mathrm{T}}b + A^{\mathrm{T}}\dot{b} - (\dot{A}^{\mathrm{T}}A + A^{\mathrm{T}}\dot{A})y_D)$
"Sketch + Diff"	$y_S = \mathrm{LLS}(A, b, S)$	$\dot{y}_S = \mathcal{J}\,\mathrm{LLS}(A, b)(y_S, \dot{A}, \dot{b}, S)$ $= \mathrm{LS}(A^{\mathrm{T}}S^{\mathrm{T}}SA, \dot{A}^{\mathrm{T}}S^{\mathrm{T}}Sb + A^{\mathrm{T}}S^{\mathrm{T}}S\dot{b}$ $- (\dot{A}^{\mathrm{T}}S^{\mathrm{T}}SA + A^{\mathrm{T}}S^{\mathrm{T}}S\dot{A})y_S)$

Table 3: Reverse mode AD Transformations.

Type	**Primal**	**Reverse Transform**
Regular	$y = \mathrm{LLS}(A, b)$	$(\bar{A}, \bar{b}) = \mathcal{J}^{\mathrm{T}}\,\mathrm{LLS}(A, b)(y, \bar{y})$ $= (-A^{\dagger\mathrm{T}}\bar{y}y^{\mathrm{T}} - Ay\bar{y}^{\mathrm{T}}M^{-1} + b\bar{y}^{\mathrm{T}}M^{-1}, A^{\dagger T}\bar{y})$
"Diff + Sketch"	$y_D = \mathrm{LLS}(A, b, S)$	$(\bar{A}, \bar{b}) = \mathcal{J}^{\mathrm{T}}\,\mathrm{LLS}(A, b)(y, \bar{y})$ $= (-AM_S^{-1}\bar{y}{y_D}^{\mathrm{T}} - Ay_D\bar{y}^{\mathrm{T}}M_S^{-1} + b\bar{y}^{\mathrm{T}}M_S^{-1}, A^{\dagger T}\bar{y})$
"Sketch + Diff"	$y_S = \mathrm{LLS}_S(A, b, S)$	$(\bar{A}_S, \bar{b}_S) = \mathcal{J}^{\mathrm{T}}\,\mathrm{LLS}_S(A, b, S)(y_S, \bar{y})$ $= (-S^{\mathrm{T}}A_S^{\dagger T}\bar{y}{y_S}^{\mathrm{T}} - S^{\mathrm{T}}SAy_S\bar{y}^{\mathrm{T}}M_S^{-1} + S^{\mathrm{T}}Sb\bar{y}^{\mathrm{T}}M_S^{-1}, S^{\mathrm{T}}A_S^{\dagger T}\bar{y})$

A.3 "Sketch and Differentiate"

Lemma 2. *The reverse mode approximation error for the term $\bar{b}$ when we approximate it by sketching matrix S can be bounded with probability $1-\delta$ as follows: $\|\bar{b}-\bar{b}_S\|_2 \leq \|\Sigma^{-1}\|_2\|\bar{y}\|_2(\epsilon+(1+\epsilon)\|I-S^{\mathrm{T}}S\|_2)$.*

Proof. Let us use Lemma 1 and sub-multiplicativity to obtain the following:

$$\begin{aligned}
\|\bar{b}-\bar{b}_S\|_2 &= \|AM^{-\mathrm{T}}\bar{y} - S^{\mathrm{T}}SAM_S{}^{-\mathrm{T}}\bar{y}_S\|_2 \\
&= \|AM^{-1}\bar{y} - AM_S^{-1}\bar{y} + AM_S^{-1}\bar{y} - S^{\mathrm{T}}SAM_S^{-1}\bar{y}\|_2 \\
&\leq \|AM^{-1} - AM_S^{-1}\|_2\|\bar{y}\|_2 + \|I-S^{\mathrm{T}}S\|_2\|AM_S^{-1}\|_2\|\bar{y}\|_2 \\
&\leq \epsilon\|\Sigma^{-1}\|_2\|\bar{y}\|_2 + \|I-S^{\mathrm{T}}S\|_2\|AM_S^{-1}\|_2\|\bar{y}\|_2 \\
&\leq \|\Sigma^{-1}\|_2\|\bar{y}\|_2(\epsilon+(1+\epsilon)\|I-S^{\mathrm{T}}S\|_2) \qquad (11)
\end{aligned}$$

□

where we used a lemma from the main paper. So, the error can be large ($\|I-S^{\mathrm{T}}S\|_2$).

Lemma 3. *The reverse mode approximation error for the term $\bar{A}$ when we approximate it using the sketching matrix S can be bounded with probability $1-\delta$.*

Proof.

$$\begin{aligned}
\|\bar{A}-\bar{A}_S\|_F &= \|-2AM^{-\mathrm{T}}\bar{y}y^{\mathrm{T}} + b\bar{y}^{\mathrm{T}}M^{-1} - (-2S^{\mathrm{T}}SAM_S{}^{-\mathrm{T}}\bar{y}_S y_S{}^{\mathrm{T}} + S^{\mathrm{T}}Sb\bar{y}_S{}^{\mathrm{T}}M_S^{-1})\|_F \\
&\leq \|2AM^{-1}\bar{y}y^{\mathrm{T}} - 2S^{\mathrm{T}}SAM_S^{-1}\bar{y}y_S{}^{\mathrm{T}}\|_F + \|b\bar{y}^{\mathrm{T}}M^{-1} - S^{\mathrm{T}}Sb\bar{y}^{\mathrm{T}}M_S^{-1})\|_F \qquad (12)
\end{aligned}$$

$$\begin{aligned}
\|AM^{-1}\bar{y}y^{\mathrm{T}} - S^{\mathrm{T}}SAM_S^{-1}\bar{y}y_S{}^{\mathrm{T}}\|_F &\leq \|AM^{-1}\bar{y}y^{\mathrm{T}} - AM_S^{-1}\bar{y}y^{\mathrm{T}}\|_F + \|AM_S^{-1}\bar{y}y^{\mathrm{T}} - S^{\mathrm{T}}SAM_S^{-1}\bar{y}y_S{}^{\mathrm{T}}\|_F \\
&\leq \epsilon\|\Sigma^{-1}\|_F\|\bar{y}\|\|y\| + \|AM_S^{-1}\bar{y}y^{\mathrm{T}} - S^{\mathrm{T}}SAM_S^{-1}\bar{y}y_S{}^{\mathrm{T}}\|_F \qquad (13)
\end{aligned}$$

□

A.4 "Differentiate and Sketch"

Lemma 4. *The reverse mode approximation error for the term $\bar{b}$ when we sketch only the computationally expensive terms by S, with probability at least $1-\delta$, satisfies: $\|\bar{b}-\bar{b}_S\|_2 \lesssim \epsilon\|\Sigma^{-1}\|_2\|\bar{y}\|_2$.*

Proof. Let us use the sketching properties and sub-multiplicativity to obtain the following:

$$\begin{aligned}
\|\bar{b}-\bar{b}_S\|_2 &= \|AM^{-\mathrm{T}}\bar{y} - AM_S{}^{-\mathrm{T}}\bar{y}_S\|_2 \\
&\approx \|U(I-U^{\mathrm{T}}S^{\mathrm{T}}SU)\Sigma^{-1}V^{\mathrm{T}}\bar{y}\|_2 \qquad \bar{y}\approx\bar{y}_S \\
&\lesssim \epsilon\|U\|_2\|\Sigma^{-1}\|_2\|\bar{y}\|_2 \\
&\lesssim \epsilon\|\Sigma^{-1}\|_2\|\bar{y}\|_2 \qquad (14)
\end{aligned}$$

□

Lemma 5. *The reverse mode approximation error for the term $\bar{A}$ when we sketch only the computationally expensive terms by S, with probability $1-1/poly(d)$, satisfies: $\|\bar{A}-\bar{A}_S\|_2 \lesssim \epsilon\|\bar{y}\|_2(\|\Sigma^{-1}\|_2\|y\|_2 + \frac{1}{1-\epsilon}\|\Sigma^{-1}\|_2\|Ay-b\|_2\|A^{\dagger}\|_2)$.*

Proof. The approximation error can be split into 3 terms such that $\|\bar{A}-\bar{A}_S\| \leq Q_1+Q_2+Q_3$ where:

$$\begin{aligned}
Q_1 &= \|b\bar{y}^{\mathrm{T}}M^{-1} - b\bar{y}_S{}^{\mathrm{T}}M_{S'}^{-1}\|_F \\
&\leq \epsilon\|b\bar{y}^{\mathrm{T}}\|_F\|\Sigma^{-1}\|_2\|\Sigma^{-1}\|_F
\end{aligned}$$

Let us bound Q_2 as follows:

$$
\begin{aligned}
Q_2 = \|AM^{-1}\bar{y}y^{\mathrm{T}} - AM_{S'}^{-1}\bar{y}y_S{}^{\mathrm{T}}\|_F &= \|A(M^{-1} - M_{S'}^{-1})\bar{y}y^{\mathrm{T}} + AM_{S'}^{-1}\bar{y}y^{\mathrm{T}} - AM_{S'}^{-1}\bar{y}y_S{}^{\mathrm{T}}\|_F \\
&\leq \|A(M^{-1} - M_{S'}^{-1})\bar{y}y^{\mathrm{T}}\|_F + \|AM_{S'}^{-1}\bar{y}(y - y_S)^{\mathrm{T}}\|_F \\
&\leq \epsilon\|\Sigma^{-1}\|_2\|\bar{y}y^{\mathrm{T}}\|_F + \|AM_{S'}^{-1}\|_2\|\bar{y}(y - y_S)^{\mathrm{T}}\|_F \\
&\leq \epsilon\|\Sigma^{-1}\|_2\|\bar{y}\|_2\|y\|_2 + \|AM_{S'}^{-1}\|_2\|\bar{y}\|_2\|(y - y_S)\|_2 \\
&\leq \epsilon\|\bar{y}\|_2(\|\Sigma^{-1}\|_2\|y\|_2 + \|AM_{S'}^{-1}\|_2\|Ay - b\|_2\|A^{\dagger}\|_2) \\
&\leq \epsilon\|\bar{y}\|_2(\|\Sigma^{-1}\|_2\|y\|_2 + (1+\epsilon)\|\Sigma^{-1}\|_2\|Ay - b\|_2\|A^{\dagger}\|_2)
\end{aligned} \tag{15}
$$

where we used the following result Price et al. (2017):

$$
\|y - y_S\|_2 \leq \epsilon\|Ay - b\|_2\|A^{\dagger}\|_2 \tag{16}
$$

and the last term Q_3 can be bounded as:

$$
Q_3 = \|Ay\bar{y}^{\mathrm{T}}M^{-1} - Ay_S\bar{y}^{\mathrm{T}}M_S^{-1}\| = \|Ay\bar{y}^{\mathrm{T}}(M^{-1} - M_S^{-1}) + Ay\bar{y}^{\mathrm{T}}M_S^{-1} - Ay_S\bar{y}^{\mathrm{T}}M_S^{-1}\| \tag{17}
$$

$$
\begin{aligned}
&\leq \epsilon\|Ay\bar{y}^{\mathrm{T}}\| + \|A(y - y_S)\bar{y}^{\mathrm{T}}M_S^{-1}\| \\
&\leq \epsilon\|Ay\bar{y}^{\mathrm{T}}\| + \epsilon\|A\|\|Ay - b\|\|\bar{y}^{\mathrm{T}}M_S^{-1}\|
\end{aligned} \tag{18}
$$

Note that all three terms Q_1, Q_2, Q_3 are $O(\epsilon)$. □

B Experiments

We plot the performance of the two proposed approaches for obtaining forward and reverse mode AD in the case of linear regression. We generate a linear regression problem by choosing the entries of matrix A and vector b from i.i.d. $N(0,1)$ (Normal distribution with mean 0 and variance 1). The differences from the two approaches, "sketch+differentiate" and "differentiate+sketch" are shown in Figure 1.

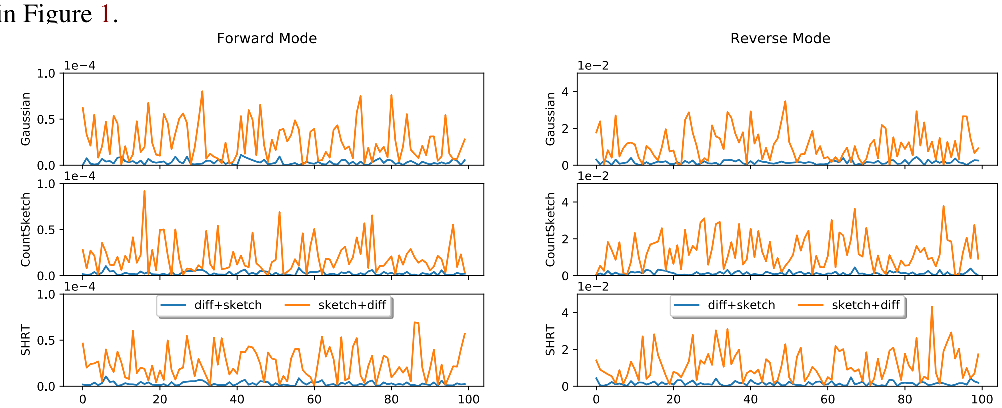

Figure 1: Numerical observation that differentiation and sketching do not commute, and that differentiation-then-sketch is more accurate. We show the forward mode along with its approximation corresponding to the three sketching matrices of Gaussian, Count-sketch and Subsampled Randomized Hadamard Transform (SRHT), on a randomly generated least squares problem of size 100000×100, along with a random perturbation. Reverse mode is shown for a subsample of 100 randomly chosen values for the variable b, where we used sign as the cost function.

References

Eric Price, Zhao Song, and David P. Woodruff. Fast regression with an l_∞ guarantee. In *44th International Colloquium on Automata, Languages, and Programming, ICALP 2017, July 10-14, 2017, Warsaw, Poland*, pages 59:1–59:14, 2017. doi: 10.4230/LIPIcs.ICALP.2017.59.