Appendix: Efficient Computation and Analysis of Distributional Shapley Values

1 Proofs

1.1 Proof of Proposition 1

Proof of Proposition 1. To this end, we fix S and γ. The ridge estimator based on S and $S \cup \{(x^*, y^*)\}$ are given by

$$\hat{\beta}_{S,\gamma} = A_{S,\gamma}^{-1} X_S^T Y_S,$$

and

$$\hat{\beta}_{S\cup\{(x^*,y^*)\},\gamma} = (X_{S\cup\{(x^*,y^*)\}}^T X_{S\cup\{(x^*,y^*)\}} + \gamma I_p)^{-1} X_{S\cup\{(x^*,y^*)\}}^T Y_{S\cup\{(x^*,y^*)\}},$$

respectively. By Sherman-Morrison formula,

$$(x^* x^{*T} + A_{S,\gamma})^{-1} = A_{S,\gamma}^{-1} - \frac{A_{S,\gamma}^{-1} x^* x^{*T} A_{S,\gamma}^{-1}}{1 + x^{*T} A_{S,\gamma}^{-1} x^*},$$

and

$$\begin{aligned}\hat{\beta}_{S\cup\{(x^*,y^*)\},\gamma} &= \hat{\beta}_{S,\gamma} + A_{S,\gamma}^{-1} x^* y^* - \frac{A_{S,\gamma}^{-1} x^* x^{*T} \hat{\beta}_{S,\gamma}}{1 + x^{*T} A_{S,\gamma}^{-1} x^*} - \frac{A_{S,\gamma}^{-1} x^* x^{*T} A_{S,\gamma}^{-1} x^* y^*}{1 + x^{*T} A_{S,\gamma}^{-1} x^*} \\ &= \hat{\beta}_{S,\gamma} + \underbrace{\frac{A_{S,\gamma}^{-1} x^* (y^* - x^{*T} \hat{\beta}_{S,\gamma})}{1 + x^{*T} A_{S,\gamma}^{-1} x^*}}_{=: f_\gamma(X_S)}.\end{aligned}$$

Since $U_{q,\gamma}(S) = (C_{\text{lin}} - \int (y - x^T \hat{\beta}_{S,\gamma})^2 dP_{X,Y}(x,y)) \mathbb{1}(|S| \geq q) = (C_{\text{lin}} - \sigma^2 - (\hat{\beta}_{S,\gamma} - \beta)^T \Sigma_X (\hat{\beta}_{S,\gamma} - \beta)) \mathbb{1}(|S| \geq q)$, for $j - 1 \geq q$, we have

$$\begin{aligned}&\mathbb{E}_{S\sim P_{X,Y}^{j-1}}[U_{q,\gamma}(S \cup \{(x^*, y^*)\})] \\ &= C_{\text{lin}} - \sigma^2 - \mathbb{E}_{S\sim P_{X,Y}^{j-1}}[(\hat{\beta}_{S\cup\{(x^*,y^*)\},\gamma} - \beta)^T \Sigma_X (\hat{\beta}_{S\cup\{(x^*,y^*)\},\gamma} - \beta)] \\ &= \mathbb{E}_{S\sim P_{X,Y}^{j-1}}[U_{q,\gamma}(S)] - \mathbb{E}_{S\sim P_{X,Y}^{j-1}}[f_\gamma(X_S)^T \Sigma_X f_\gamma(X_S)] - 2\mathbb{E}_{S\sim P_{X,Y}^{j-1}}[f_\gamma(X_S)^T \Sigma_X (\hat{\beta}_{S,\gamma} - \beta)].\end{aligned}$$

Therefore,

$$\begin{aligned}&\mathbb{E}_{S\sim P_{X,Y}^{j-1}}[U_{q,\gamma}(S \cup \{(x^*, y^*)\}) - U_{q,\gamma}(S)] \\ &= -(\mathbb{E}_{S\sim P_{X,Y}^{j-1}}[f_\gamma(X_S)^T \Sigma_X f_\gamma(X_S)] + 2\mathbb{E}_{S\sim P_{X,Y}^{j-1}}[f_\gamma(X_S)^T \Sigma_X (\hat{\beta}_{S,\gamma} - \beta)]),\end{aligned}$$

and thus for $q \geq p + 1$, DShapley is

$$\begin{aligned}&\nu((x^*, y^*); U_{q,\gamma}, P_{X,Y}, m) \\ &= \frac{1}{m} \sum_{j=1}^{m} \mathbb{E}_{S\sim P_{X,Y}^{j-1}}[U_{q,\gamma}(S \cup \{(x^*, y^*)\}) - U_{q,\gamma}(S)]\end{aligned}$$

$$= (C_{\mathrm{lin}} - \sigma^2 - \mathbb{E}_{S\sim P^{q-1}_{X,Y}}[(\hat{\beta}_{S,\gamma} - \beta)^T \Sigma_X (\hat{\beta}_{S,\gamma} - \beta)])$$
$$- \frac{1}{m}\sum_{j=q}^{m}\left(\mathbb{E}_{S\sim P^{j-1}_{X,Y}}[f_\gamma(X_S)^T \Sigma_X f_\gamma(X_S)] + 2\mathbb{E}_{S\sim P^{j-1}_{X,Y}}[f_\gamma(X_S)^T \Sigma_X (\hat{\beta}_{S,\gamma} - \beta)]\right).$$

[Step 1] Computation of $\mathbb{E}[f_\gamma(X_S)^T \Sigma_X f_\gamma(X_S) \mid X_S]$.
We set $e^*_{S,\gamma} = y^* - x^{*T}\mathbb{E}[\hat{\beta}_{S,\gamma} \mid X_S] = y^* - x^{*T} A^{-1}_{S,\gamma}(X_S^T X_S)\beta$, then

$$\mathbb{E}[f_\gamma(X_S) \mid X_S] = \frac{A^{-1}_{S,\gamma} x^* e^*_{S,\gamma}}{1 + x^{*T} A^{-1}_{S,\gamma} x^*},$$

and $\mathrm{Cov}[\hat{\beta}_{S,\gamma} \mid X_S] = A^{-1}_{S,\gamma}(X_S^T X_S)A^{-1}_{S,\gamma}\sigma^2 = A^{-1}_{S,\gamma}(A_{S,\gamma} - \gamma I_p)A^{-1}_{S,\gamma}\sigma^2 = A^{-1}_{S,\gamma}(I_p - \gamma A^{-1}_{S,\gamma})\sigma^2 = (A^{-1}_{S,\gamma} - \gamma A^{-2}_{S,\gamma})\sigma^2 =: M_{S,\gamma}\sigma^2$ gives

$$\mathrm{Cov}[f_\gamma(X_S) \mid X_S] = \frac{A^{-1}_{S,\gamma} x^* x^{*T} M_{S,\gamma} x^* x^{*T} A^{-1}_{S,\gamma}}{(1 + x^{*T} A^{-1}_{S,\gamma} x^*)^2}\sigma^2.$$

Thus,

$$\mathbb{E}[f_\gamma(X_S)^T \Sigma_X f_\gamma(X_S) \mid X_S] = \frac{x^{*T} A^{-1}_{S,\gamma} \Sigma_X A^{-1}_{S,\gamma} x^*}{(1 + x^{*T} A^{-1}_{S,\gamma} x^*)^2} e^{*2}_{S,\gamma} + \frac{x^{*T} A^{-1}_{S,\gamma} \Sigma_X A^{-1}_{S,\gamma} x^*}{(1 + x^{*T} A^{-1}_{S,\gamma} x^*)^2} x^{*T} M_{S,\gamma} x^* \sigma^2.$$

Since

$$e^*_{S,\gamma} = e^* + x^{*T}(\beta - A^{-1}_{S,\gamma}(X_S^T X_S)\beta) = e^* + \gamma x^{*T} A^{-1}_{S,\gamma}\beta,$$

and $M_{S,\gamma} = A^{-1}_{S,\gamma} - \gamma A^{-2}_{S,\gamma}$, we have

$$\mathbb{E}[f_\gamma(X_S)^T \Sigma_X f_\gamma(X_S) \mid X_S] = \frac{x^{*T} A^{-1}_{S,\gamma} \Sigma_X A^{-1}_{S,\gamma} x^*}{(1 + x^{*T} A^{-1}_{S,\gamma} x^*)^2}\left(\frac{e^{*2}}{\sigma^2} + x^{*T} A^{-1}_{S,\gamma} x^*\right)\sigma^2 + h_1(\gamma),$$

where $h_1(\gamma)$ is some explicit term such that $\lim_{\gamma\to 0+} h_1(\gamma)/(\gamma\log(\gamma))$ and $h_1(0) = 0$.

[Step 2] Computation of $\mathbb{E}[f_\gamma(X_S)^T \Sigma_X (\hat{\beta}_{S,\gamma} - \beta) \mid X_S]$.

$$\begin{aligned}
&\mathbb{E}[f_\gamma(X_S)^T \Sigma_X (\hat{\beta}_{S,\gamma} - \beta) \mid X_S] \\
&= \mathbb{E}\Big[\frac{(y^* - x^{*T}\hat{\beta}_{S,\gamma}) x^{*T} A^{-1}_{S,\gamma} \Sigma_X (\hat{\beta}_{S,\gamma} - \beta)}{1 + x^{*T} A^{-1}_{S,\gamma} x^*} \mid X_S\Big] \\
&= -\gamma \frac{e^* x^{*T} A^{-1}_{S,\gamma} \Sigma_X A^{-1}_{S,\gamma}\beta}{1 + x^{*T} A^{-1}_{S,\gamma} x^*} - \mathbb{E}_S\Big[\frac{(\hat{\beta}_{S,\gamma} - \beta)^T x^* x^{*T} A^{-1}_{S,\gamma} \Sigma_X (\hat{\beta}_{S,\gamma} - \beta)}{1 + x^{*T} A^{-1}_{S,\gamma} x^*} \mid X_S\Big] \\
&= -\gamma \frac{e^* x^{*T} A^{-1}_{S,\gamma} \Sigma_X A^{-1}_{S,\gamma}\beta}{1 + x^{*T} A^{-1}_{S,\gamma} x^*} - \gamma^2 \frac{\beta^T A^{-1}_{S,\gamma} x^* x^{*T} A^{-1}_{S,\gamma} \Sigma_X A^{-1}_{S,\gamma}\beta}{1 + x^{*T} A^{-1}_{S,\gamma} x^*} - \frac{x^{*T} A^{-1}_{S,\gamma} \Sigma_X M_{S,\gamma} x^*}{1 + x^{*T} A^{-1}_{S,\gamma} x^*}\sigma^2 \\
&= -\frac{x^{*T} A^{-1}_{S,\gamma} \Sigma_X A^{-1}_{S,\gamma} x^*}{1 + x^{*T} A^{-1}_{S,\gamma} x^*}\sigma^2 + h_2(\gamma).
\end{aligned}$$

where $h_2(\gamma)$ is some explicit term such that $\lim_{\gamma\to 0+} h_2(\gamma)/(\gamma\log(\gamma)) = 0$ and $h_2(0) = 0$.

Hence, by setting $C_{\mathrm{lin}} = \sigma^2 + \mathbb{E}_{S\sim P^{q-1}_{X,Y}}[(\hat{\beta}_{S,\gamma} - \beta)^T \Sigma_X (\hat{\beta}_{S,\gamma} - \beta)]$, we have

$$\nu((x^*, y^*); U_{q,\gamma}, P_{X,Y}, m)$$

$$
\begin{aligned}
&= C_{\text{lin}} - \sigma^2 - \mathbb{E}_{S\sim P_{X,Y}^{q-1}}[(\hat{\beta}_{S,\gamma}-\beta)^T\Sigma_X(\hat{\beta}_{S,\gamma}-\beta)] \\
&- \frac{1}{m}\sum_{j=q}^{m}\mathbb{E}_{X_S\sim P_X^{j-1}}\left[\frac{x^{*T}A_{S,\gamma}^{-1}\Sigma_X A_{S,\gamma}^{-1}x^*}{(1+x^{*T}A_{S,\gamma}^{-1}x^*)^2}(e^{*2}-(2+x^{*T}A_{S,\gamma}^{-1}x^*)\sigma^2)\right]+h(\gamma), \qquad (9)\\
&= -\frac{1}{m}\sum_{j=q}^{m}\mathbb{E}_{X_S\sim P_X^{j-1}}\left[\frac{x^{*T}A_{S,\gamma}^{-1}\Sigma_X A_{S,\gamma}^{-1}x^*}{(1+x^{*T}A_{S,\gamma}^{-1}x^*)^2}(e^{*2}-(2+x^{*T}A_{S,\gamma}^{-1}x^*)\sigma^2)\right]+h(\gamma),\\
&= \frac{1}{m}\sum_{j=q}^{m}\mathbb{E}_{X_S\sim P_X^{j-1}}\left[\frac{x^{*T}A_{S,\gamma}^{-1}\Sigma_X A_{S,\gamma}^{-1}x^*}{(1+x^{*T}A_{S,\gamma}^{-1}x^*)^2}((2+x^{*T}A_{S,\gamma}^{-1}x^*)\sigma^2-e^{*2})\right]+h(\gamma),
\end{aligned}
$$

for some $h(\gamma)$ such that $\lim_{\gamma\to 0+} h(\gamma)/(\gamma\log(\gamma))=0$ and $h(0)=0$. □

1.2 Proof of Theorem 2

Proof of Theorem 2. By plugging $\gamma=0$ into Equation (9), for $q\geq p+3$, DShapley is given by

$$
\begin{aligned}
&\nu((x^*,y^*);U_{q,0},P_{X,Y},m)\\
&= C_{\text{lin}} - \sigma^2 - \mathbb{E}_{S\sim P_{X,Y}^{q-1}}[(\hat{\beta}_{S,\gamma}-\beta)^T\Sigma_X(\hat{\beta}_{S,\gamma}-\beta)]\\
&+ \frac{\sigma^2}{m}\sum_{j=q}^{m}\left((1-\frac{e^{*2}}{\sigma^2})\mathbb{E}_{X_S\sim P_X^{j-1}}\left[\frac{\tilde{x}^{*T}(\tilde{X}_S^T\tilde{X}_S)^{-2}\tilde{x}^*}{(1+\tilde{x}^{*T}(\tilde{X}_S^T\tilde{X}_S)^{-1}\tilde{x}^*)^2}\right]+\mathbb{E}_{X_S\sim P_X^{j-1}}\left[\frac{\tilde{x}^{*T}(\tilde{X}_S^T\tilde{X}_S)^{-2}\tilde{x}^*}{1+\tilde{x}^{*T}(\tilde{X}_S^T\tilde{X}_S)^{-1}\tilde{x}^*}\right]\right), \qquad (10)
\end{aligned}
$$

where $\tilde{X}_S=X_S\Sigma_X^{-1/2}$ and $\tilde{x}^*=\Sigma_X^{-1/2}x^*$, *i.e.*, a normalized version. Note that $(\tilde{X}_S^T\tilde{X}_S)^{-1}$ follows an inverse-Wishart distribution and its mean is $I_p/(q-1-p-1)$. Therefore,

$$
-\frac{\sigma^2}{m}\text{tr}(\mathbb{E}_{X_S\sim P_X^{q-1}}[(\tilde{X}_S^T\tilde{X}_S)^{-1}])=-\frac{\sigma^2}{m}\frac{p}{q-p-2}.
$$

Now it is enough to compute the following expectations:

$$
\mathbb{E}_{X_S\sim P_X^{j-1}}[\frac{\tilde{x}^{*T}(\tilde{X}_S^T\tilde{X}_S)^{-2}\tilde{x}^*}{(1+\tilde{x}^{*T}(\tilde{X}_S^T\tilde{X}_S)^{-1}\tilde{x}^*)^2}] \quad\text{and}\quad \mathbb{E}_{X_S\sim P_X^{j-1}}[\frac{\tilde{x}^{*T}(\tilde{X}_S^T\tilde{X}_S)^{-2}\tilde{x}^*}{1+\tilde{x}^{*T}(\tilde{X}_S^T\tilde{X}_S)^{-1}\tilde{x}^*}].
$$

[Step 1] For any $p\times p$ orthogonal matrix Γ, we have $\Gamma(\tilde{X}_S^T\tilde{X}_S)\Gamma^T\sim W_p(|S|,I_p)$ due to $\tilde{X}_S^T\tilde{X}_S\sim W_p(|S|,I_p)$. We choose an orthogonal matrix Γ with the first column is $(\tilde{x}^{*T}\tilde{x}^*)^{-1/2}\tilde{x}^*$ and let $V:=\Gamma(\tilde{X}_S^T\tilde{X}_S)\Gamma^T$. Then, $\tilde{x}^{*T}(\tilde{X}_S^T\tilde{X}_S)^{-1}\tilde{x}^*=(\Gamma\tilde{x}^*)^TV^{-1}(\Gamma\tilde{x}^*)=\tilde{x}^{*T}\tilde{x}^*v^{11}$ where $V^{-1}=(v^{ij})$. Similarly, we obtain $\tilde{x}^{*T}(\tilde{X}_S^T\tilde{X}_S)^{-2}\tilde{x}^*=\tilde{x}^{*T}\tilde{x}^*\sum_{j=1}^{p}(v^{1j})^2$.

Now we let $V=TT^T$ where T is an upper triangular matrix with positive diagonal elements as

$$
T=\begin{pmatrix} t_{11} & \mathbf{t}^T \\ 0 & T_{22}\end{pmatrix}.
$$

Then,

$$
T^{-1}=\begin{pmatrix} t_{11}^{-1} & -t_{11}^{-1}\mathbf{t}^TT_{22}^{-1} \\ 0 & T_{22}^{-1}\end{pmatrix},\quad V^{-1}=\begin{pmatrix} t_{11}^{-2} & -t_{11}^{-2}\mathbf{t}^TT_{22}^{-1} \\ -t_{11}^{-2}(T_{22}^T)^{-1}\mathbf{t} & (T_{22}T_{22}^T)^{-1}+t_{11}^{-2}(T_{22}^T)^{-1}\mathbf{t}\mathbf{t}^TT_{22}^{-1}\end{pmatrix}.
$$

Therefore,

$$
\frac{\tilde{x}^{*T}(\tilde{X}_S^T\tilde{X}_S)^{-2}\tilde{x}^*}{(1+\tilde{x}^{*T}(\tilde{X}_S^T\tilde{X}_S)^{-1}\tilde{x}^*)^2}=\frac{\tilde{x}^{*T}\tilde{x}^*(t_{11}^{-4}+t_{11}^{-4}\mathbf{t}^T(T_{22}^TT_{22})^{-1}\mathbf{t})}{(1+\tilde{x}^{*T}\tilde{x}^*t_{11}^{-2})^2}=\frac{\tilde{x}^{*T}\tilde{x}^*(1+\mathbf{t}^T(T_{22}^TT_{22})^{-1}\mathbf{t})}{(\tilde{x}^{*T}\tilde{x}^*+t_{11}^2)^2}.
$$

Due to Gupta and Nagar (1999, Theorem 3.3.5), t_{11}^2 is independent to $\mathbf{t}^T(T_{22}^TT_{22})^{-1}\mathbf{t}$ with $t_{11}^2\sim\chi_{|S|-p+1}^2$. Furthermore, by Gupta and Nagar (1999, Theorem 3.3.28), $\mathbf{t}^T(T_{22}^TT_{22})^{-1}\mathbf{t}\sim\frac{p-1}{|S|-p+2}F_{p-1,|S|-p+2}$. That is,

$$
\mathbb{E}_{X_S\sim P_X^{j-1}}[\frac{\tilde{x}^{*T}(\tilde{X}_S^T\tilde{X}_S)^{-2}\tilde{x}^*}{(1+\tilde{x}^{*T}(\tilde{X}_S^T\tilde{X}_S)^{-1}\tilde{x}^*)^2}]=\tilde{x}^{*T}\tilde{x}^*\mathbb{E}[(1+\mathbf{t}^T(T_{22}^TT_{22})^{-1}\mathbf{t})]\mathbb{E}[\frac{1}{(\tilde{x}^{*T}\tilde{x}^*+t_{11}^2)^2}]
$$

$$= \tilde{x}^{*T}\tilde{x}^* \frac{|S|-1}{|S|-p}\mathbb{E}[\frac{1}{(\tilde{x}^{*T}\tilde{x}^* + t_{11}^2)^2}].$$

[Step 2] Similarly, we have

$$\frac{\tilde{x}^{*T}(\tilde{X}_S^T\tilde{X}_S)^{-2}\tilde{x}^*}{1+\tilde{x}^{*T}(\tilde{X}_S^T\tilde{X}_S)^{-1}\tilde{x}^*} = \frac{\tilde{x}^{*T}\tilde{x}^*(t_{11}^{-4} + t_{11}^{-4}\mathbf{t}^T(T_{22}^TT_{22})^{-1}\mathbf{t})}{1+\tilde{x}^{*T}\tilde{x}^*t_{11}^{-2}}$$
$$= (1+\mathbf{t}^T(T_{22}^TT_{22})^{-1}\mathbf{t})\left(\frac{1}{t_{11}^2} - \frac{1}{t_{11}^2+\tilde{x}^{*T}\tilde{x}^*}\right),$$

and

$$\mathbb{E}_{X_S\sim P_X^{j-1}}[\frac{\tilde{x}^{*T}(\tilde{X}_S^T\tilde{X}_S)^{-2}\tilde{x}^*}{1+\tilde{x}^{*T}(\tilde{X}_S^T\tilde{X}_S)^{-1}\tilde{x}^*}] = \frac{|S|-1}{|S|-p}\mathbb{E}[\left(\frac{1}{t_{11}^2} - \frac{1}{t_{11}^2+\tilde{x}^{*T}\tilde{x}^*}\right)]$$
$$= \frac{|S|-1}{|S|-p}\left(\frac{1}{|S|-p-1} - \mathbb{E}[\frac{1}{t_{11}^2+\tilde{x}^{*T}\tilde{x}^*}]\right).$$

[Step 3] Therefore, for any $q \geq p+3$ and Chi-squared distributions $T_j \sim \chi^2_{j-p+1}$ (or equivalently Gamma distributions $T_j \sim \text{Gamma}((j-p+1)/2, 1/2)$), we have

$$\begin{aligned}
&\nu((x^*,y^*);U_{q,0},P_{X,Y},m)\\
&= C_{\text{lin}} - \sigma^2 - \mathbb{E}_{S\sim P_{X,Y}^{q-1}}[(\hat{\beta}_{S,\gamma}-\beta)^T\Sigma_X(\hat{\beta}_{S,\gamma}-\beta)]\\
&+ \frac{\sigma^2}{m}\sum_{j=q}^{m}\left((1-\frac{e^{*2}}{\sigma^2})\frac{j-1}{j-p}\mathbb{E}[\frac{\tilde{x}^{*T}\tilde{x}^*}{(\tilde{x}^{*T}\tilde{x}^*+T_j)^2}] + \frac{j-1}{j-p}\left(\frac{1}{j-p-1} - \mathbb{E}[\frac{1}{\tilde{x}^{*T}\tilde{x}^*+T_j}]\right)\right)
\end{aligned}$$

By setting

$$C_{\text{lin}} = \sigma^2 + \mathbb{E}_{S\sim P_{X,Y}^{q-1}}[(\hat{\beta}_{S,\gamma}-\beta)^T\Sigma_X(\hat{\beta}_{S,\gamma}-\beta)] - \frac{\sigma^2}{m}\sum_{j=q}^{m}\frac{j-1}{j-p}\frac{1}{j-p-1},$$

we have

$$\begin{aligned}
&\nu((x^*,y^*);U_{q,0},P_{X,Y},m)\\
&= \frac{\sigma^2}{m}\sum_{j=q}^{m}\left((1-\frac{e^{*2}}{\sigma^2})\frac{j-1}{j-p}\mathbb{E}[\frac{\tilde{x}^{*T}\tilde{x}^*}{(\tilde{x}^{*T}\tilde{x}^*+T_j)^2}] - \frac{j-1}{j-p}\mathbb{E}[\frac{1}{\tilde{x}^{*T}\tilde{x}^*+T_j}]\right)\\
&= -\frac{1}{m}\sum_{j=q}^{m}\mathbb{E}\left[\frac{j-1}{j-p}\frac{\left(x^{*T}\Sigma_X^{-1}x^*e^{*2}+T_j\sigma^2\right)}{(x^{*T}\Sigma_X^{-1}x^*+T_j)^2}\right].
\end{aligned}$$

□

1.3 Proof of Theorem 3

To begin, we first define some notations and a useful lemma. Let $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ be the smallest and largest singular values of a matrix A. For a sub-Gaussian random variable X, we denote its sub-Gaussian norm by $\|X\|_{\psi_2} := \sup_{p\geq 1} p^{-1/2}(\mathbb{E}|X|^p)^{1/p}$. For a sub-Gaussian random vector X, we denote its sub-Gaussian norm by $\|X\|_{\psi_2} := \sup_{x^Tx=1}\left\|\langle X,x\rangle\right\|_{\psi_2}$. Lastly, we quote the non-asymptotic eigenvalue bounds by Vershynin (2010, Theorem 5.39).

Lemma 6. *Suppose that $\tilde{X}_S$ is a matrix whose rows are independent sub-Gaussian isotropic random vectors in $\mathbb{R}^p$, then for every $t \geq 0$, with probability at least $1-2\exp(-ct^2)$ one has*

$$\sqrt{|S|}(1-\delta_{|S|}) = \sqrt{|S|} - C\sqrt{p} - t \leq \lambda_{\min}(\tilde{X}_S) \leq \lambda_{\max}(\tilde{X}_S) \leq \sqrt{|S|} + C\sqrt{p} + t = \sqrt{|S|}(1+\delta_{|S|}),$$

where $\delta_{|S|} = (C\sqrt{p}+t)/\sqrt{|S|}$ and C, c are two constants depending only on the sub-Gaussian norm.

Proof of Theorem 3. [Step 1] We provide a proof for the upper bound only, but the similar procedure can show the lower bound. To this end, we fix S and let $\tilde{X}_S = X_S \Sigma_X^{-1/2}$, $\tilde{x}^* = \Sigma_X^{-1/2} x^*$, and $\tilde{A}_\gamma = (\tilde{X}_S^T \tilde{X}_S + \gamma \Sigma_X^{-1})$. Then, we have

$$\frac{x^{*T} A_\gamma^{-1} \Sigma_X A_\gamma^{-1} x^*}{1 + x^{*T} A_\gamma^{-1} x^*} \frac{(2 + x^{*T} A_\gamma^{-1} x^*)\sigma^2 - e^{*2}}{1 + x^{*T} A_\gamma^{-1} x^*} = \frac{(\tilde{x}^{*T} \tilde{A}_\gamma^{-2} \tilde{x}^*)(2 + \tilde{x}^{*T} \tilde{A}_\gamma^{-1} \tilde{x}^*)\sigma^2 - e^{*2}}{(1 + \tilde{x}^{*T} \tilde{A}_\gamma^{-1} \tilde{x}^*)^2}. \tag{11}$$

Due to $\lambda_{\max}(AB) \leq \lambda_{\max}(A)\lambda_{\max}(B)$, we have

$$\begin{aligned}
&\frac{(\tilde{x}^{*T} \tilde{A}_\gamma^{-2} \tilde{x}^*)(2 + \tilde{x}^{*T} \tilde{A}_\gamma^{-1} \tilde{x}^*)\sigma^2 - e^{*2}}{(1 + \tilde{x}^{*T} \tilde{A}_\gamma^{-1} \tilde{x}^*)^2} \\
&\leq \frac{\tilde{x}^{*T} \tilde{x}^* \lambda_{\max}(\tilde{A}_\gamma^{-2})(2 + \tilde{x}^{*T} \tilde{x}^* \lambda_{\max}(\tilde{A}_\gamma^{-1}))}{(1 + \tilde{x}^{*T} \tilde{x}^* \lambda_{\min}(\tilde{A}_\gamma^{-1}))^2} \sigma^2 - \frac{1}{(1 + \tilde{x}^{*T} \tilde{x}^* \lambda_{\max}(\tilde{A}_\gamma^{-1}))^2} e^{*2}.
\end{aligned}$$

Since $|y_i| \leq B_{\mathcal{Y}}$ and $\hat{\beta}_S^{\mathrm{R}} = \mathrm{argmin}_\beta (Y_S - X_S\beta)^T (Y_S - X_S\beta) + \gamma \|\beta\|_2^2$, we obtain boundedness of $\left\|\hat{\beta}_S^{\mathrm{R}}\right\|_2^2$, *i.e.*, $\left\|\hat{\beta}_S^{\mathrm{R}}\right\|_2^2 \leq \gamma^{-1} Y_S^T Y_S \leq \gamma^{-1} m B_{\mathcal{Y}}^2$ for any $S \subseteq \mathcal{X} \times \mathcal{Y}$. That means, $U_q^{\mathrm{R}}(S)$ is bounded, and thus Equation (11) is bounded as well. Let say the bound is C_{bdd}.

[Step 2] Using Lemma 6 with $t_{|S|} = \sqrt{\frac{\log(|S|m^{1/2})}{c}}$, the following holds with probability at least $1 - 2/(|S|m^{1/2})$.

$$\sqrt{|S|}(1 - \delta_{|S|}) = \sqrt{|S|} - C\sqrt{p} - t \leq \lambda_{\min}(\tilde{X}_S) \leq \lambda_{\max}(\tilde{X}_S) \leq \sqrt{|S|} + C\sqrt{p} + t = \sqrt{|S|}(1 + \delta_{|S|}),$$

where $\delta_{|S|} = (C\sqrt{p} + \sqrt{\frac{\log(|S|m)}{2c}})/\sqrt{|S|}$. We denote the set where the inequalities hold by $\Omega_{|S|}$ and we obtain the following bounds.

$$\begin{aligned}
&\mathbb{E}_{X_S \sim P_X^{j-1}} \left[\frac{(\tilde{x}^{*T} \tilde{A}_\gamma^{-2} \tilde{x}^*)(2 + \tilde{x}^{*T} \tilde{A}_\gamma^{-1} \tilde{x}^*)\sigma^2 - e^{*2}}{(1 + \tilde{x}^{*T} \tilde{A}_\gamma^{-1} \tilde{x}^*)^2} \right] \\
&\leq \int_{\Omega_{|S|}} \frac{\tilde{x}^{*T} \tilde{x}^* \lambda_{\max}(\tilde{A}_\gamma^{-2})(2 + \tilde{x}^{*T} \tilde{x}^* \lambda_{\max}(\tilde{A}_\gamma^{-1}))}{(1 + \tilde{x}^{*T} \tilde{x}^* \lambda_{\min}(\tilde{A}_\gamma^{-1}))^2} \sigma^2 dP - \int_{\Omega_{|S|}} \frac{1}{(1 + \tilde{x}^{*T} \tilde{x}^* \lambda_{\max}(\tilde{A}_\gamma^{-1}))^2} e^{*2} dP \\
&+ \int_{\Omega^c} C_{\mathrm{bdd}} dP \\
&\leq \frac{\tilde{x}^{*T} \tilde{x}^* (|S|(1 - \delta_{|S|})^2 + \gamma \lambda_{\min}(\Sigma_X^{-1}))^{-2}}{(1 + \tilde{x}^{*T} \tilde{x}^* (|S|(1 + \delta_{|S|})^2 + \gamma \lambda_{\max}(\Sigma_X^{-1}))^{-1})^2} \left(2 + \tilde{x}^{*T} \tilde{x}^* (|S|(1 - \delta_{|S|})^2 + \gamma \lambda_{\min}(\Sigma_X^{-1}))^{-1}\right) \sigma^2 \\
&- \frac{e^{*2}}{(1 + \tilde{x}^{*T} \tilde{x}^* (|S|(1 - \delta_{|S|})^2 + \gamma \lambda_{\min}(\Sigma_X^{-1}))^{-1})^2} + C_{\mathrm{bdd}} P(\Omega_{|S|}^c),
\end{aligned}$$

where the second inequality is due to $\lambda_{\min}(A + B) \geq \lambda_{\min}(A) + \lambda_{\min}(B)$ and $\lambda_{\max}(A + B) \leq \lambda_{\max}(A) + \lambda_{\max}(B)$. Hence,

$$\begin{aligned}
&\nu((x^*, y^*); U_{q,\gamma}, P_{X,Y}, m) \\
&\leq \frac{1}{m} \sum_{j=q-1}^{m-1} \frac{\tilde{x}^{*T} \tilde{x}^* (j(1 - \delta_j)^2 + \gamma \lambda_{\min}(\Sigma_X^{-1}))^{-2}}{(1 + \tilde{x}^{*T} \tilde{x}^* (j(1 + \delta_j)^2 + \gamma \lambda_{\max}(\Sigma_X^{-1}))^{-1})^2} \left(2 + \tilde{x}^{*T} \tilde{x}^* (j(1 - \delta_j)^2 + \gamma \lambda_{\min}(\Sigma_X^{-1}))^{-1}\right) \sigma^2 \\
&- \frac{1}{m} \sum_{j=q-1}^{m-1} \frac{e^{*2}}{(1 + \tilde{x}^{*T} \tilde{x}^* (j(1 - \delta_j)^2 + \gamma \lambda_{\min}(\Sigma_X^{-1}))^{-1})^2} + \frac{C_{\mathrm{bdd}}}{m} \sum_{j=q-1}^{m-1} P(\Omega_j^c) + h(\gamma) \\
&= \frac{1}{m} \sum_{j=q-1}^{m-1} \frac{\tilde{x}^{*T} \tilde{x}^* \Lambda_{\mathrm{upper}}^2(j)}{(1 + \tilde{x}^{*T} \tilde{x}^* \Lambda_{\mathrm{lower}}(j))^2} \left(2 + \tilde{x}^{*T} \tilde{x}^* \Lambda_{\mathrm{upper}}(j)\right) \sigma^2 \\
&- \frac{1}{m} \sum_{j=q-1}^{m-1} \frac{e^{*2}}{(1 + \tilde{x}^{*T} \tilde{x}^* \Lambda_{\mathrm{upper}}(j))^2} + \frac{C_{\mathrm{bdd}}}{m} \sum_{j=q-1}^{m-1} P(\Omega_j^c) + h(\gamma),
\end{aligned}$$

where $\Lambda_{\text{upper}}(j) := (j(1-\delta_j)^2 + \gamma\lambda_{\min}(\Sigma_X^{-1}))^{-1}$ and $\Lambda_{\text{lower}}(j) := (j(1+\delta_j)^2 + \gamma\lambda_{\max}(\Sigma_X^{-1}))^{-1}$ for $j \in \mathbb{N}$. Lastly, $\frac{1}{m}\sum_{j=q-1}^{m-1} P(\Omega_j^c) = \frac{1}{m}\sum_{j=q-1}^{m-1} \frac{2}{j\sqrt{m}} \leq 4\frac{\log(m)}{m^{3/2}}$ concludes a proof. □

Remark 1. *It is noteworthy that the eigenvalues of $A_{S,\gamma}^{-1}$ are contained in $[\Lambda_{\text{lower}}(j)$, and $\Lambda_{\text{upper}}(j)]$ with high probability. By Lemma 6, on Ω_j, we have*

$$j(1-\delta_j)^2 + \gamma\lambda_{\min}(\Sigma_X^{-1}) \leq \lambda_{\min}(A_{S,\gamma}) \leq \lambda_{\max}(A_{S,\gamma}) \leq j(1+\delta_j)^2 + \gamma\lambda_{\max}(\Sigma_X^{-1}),$$

and thus

$$\Lambda_{\text{lower}}(j) \leq \lambda_{\min}(A_{S,\gamma}^{-1}) \leq \lambda_{\max}(A_{S,\gamma}^{-1}) \leq \Lambda_{\text{upper}}(j).$$

1.4 Proof of Corollary 4

We first provide a detailed version of Corollary 4.

Corollary 7 (DShapley in binary classification; a detailed version)**.** *Assume $\mathbb{E}[Y \mid X] = \text{logit}^{-1}(X^T\beta)$ and X are sub-Gaussian in $\mathbb{R}^p$ with $\mathbb{E}(XX^T) = \Sigma_X$. For a point (x^*, y^*), let $\pi^* = \text{logit}^{-1}(x^{*T}\beta)$, $w^* = \pi^*(1-\pi^*)$, and $z^* = x^{*T}\beta + (y^* - \pi^*)/w^*$. Then, for any $q \geq p+3$ and some fixed constant C_{lin}, DShapley of a point $\left((w^*)^{1/2}x^*, (w^*)^{1/2}z^*\right)$ has the following upper and lower bounds.*

$$\begin{aligned}
&\frac{1}{m}\sum_{j=q-1}^{m-1} \frac{w^* x^{*T}\tilde{\Sigma}_X^{-1}x^*\tilde{\Lambda}_{\text{lower}}^2(j)}{(1+w^*x^{*T}\tilde{\Sigma}_X^{-1}x^*\tilde{\Lambda}_{\text{upper}}(j))^2}\left((2+w^*x^{*T}\tilde{\Sigma}_X^{-1}x^*\tilde{\Lambda}_{\text{lower}}(j)) - \tilde{\Lambda}_{\text{ratio}}^{-1}(j)e_{\text{b}}^{*2}\right)\\
&\leq \nu\left(\left((w^*)^{1/2}x^*, (w^*)^{1/2}z^*\right); U_{q,0}, P_{\tilde{X},\tilde{Z}}, m\right) + o\left(\frac{1}{m}\right)\\
&\leq \frac{1}{m}\sum_{j=q-1}^{m-1} \frac{w^* x^{*T}\tilde{\Sigma}_X^{-1}x^*\tilde{\Lambda}_{\text{upper}}^2(j)}{(1+w^*x^{*T}\tilde{\Sigma}_X^{-1}x^*\tilde{\Lambda}_{\text{lower}}(j))^2}\left((2+w^*x^{*T}\tilde{\Sigma}_X^{-1}x^*\tilde{\Lambda}_{\text{upper}}(j)) - \tilde{\Lambda}_{\text{ratio}}(j)e_{\text{b}}^{*2}\right),
\end{aligned}$$

where $e_{\text{b}}^{*2} := (w^*)^{-1}(y^*-\pi^*)^2$, the function h is defined in Proposition 1, $\tilde{\Sigma}_X := \mathbb{E}[wXX^T]$, $\tilde{\Lambda}_{\text{upper}}(j) := (j(1-\delta_j)^2)^{-1}$, $\tilde{\Lambda}_{\text{lower}}(j) := (j(1+\delta_j)^2)^{-1}$, and $\delta_j = (C\sqrt{p} + \sqrt{\frac{\log(jm)}{2c}})/\sqrt{j}$ for $j \in \mathbb{N}$ and certain constants c, C as in the proof of Theorem 3. Lastly,

$$\tilde{\Lambda}_{\text{ratio}}(j) = \left(\frac{1+w^*x^{*T}\tilde{\Sigma}_X^{-1}x^*\tilde{\Lambda}_{\text{lower}}(j)}{1+w^*x^{*T}\tilde{\Sigma}_X^{-1}x^*\tilde{\Lambda}_{\text{upper}}(j)}\right)^2.$$

Proof of Corollary 4. Since $\mathbb{E}[Y \mid X] = \text{logit}^{-1}(X^T\beta) = \pi$, we have $\mathbb{E}[w^{1/2}Z \mid w^{1/2}X] = w^{1/2}X^T\beta$ and $\text{Var}[w^{1/2}Z \mid w^{1/2}X] = 1$. Furthermore, by the definition of sub-Gaussian, $w^{1/2}X$ is also sub-Gaussian with $\tilde{\Sigma}_X$ because $w \leq 1$ and X are sub-Gaussian with $\mathbb{E}(XX^T) = \Sigma_X$. With the notations, Theorem 3 with $\gamma = 0$ gives the upper and lower bounds. □

1.5 Proof of Theorem 5

Proof of Theorem 5. Let $S^* = \{z_1^*, \ldots, z_n^*\}$. A simple algebra gives $\hat{p}_{S\cup S^*}(z) = \frac{1}{|S|+n}(\sum_{j=1}^n k(z, z_j^*) + |S|\hat{p}_S(z)) = \frac{1}{|S|+n}\sum_{j=1}^n k(z, z_j^*) + \frac{|S|}{|S|+n}\hat{p}_S(z) = \hat{p}_S(z) + \frac{n}{|S|+n}(\frac{1}{n}\sum_{j=1}^n k(z, z_j^*) - \hat{p}_S(z))$. Note that $\frac{1}{n}\sum_{j=1}^n k(z, z_j^*) = \hat{p}_{S^*}(z)$. For $|S| \geq 1$, we have

$$\begin{aligned}
&U(S\cup S^*) - U(S)\\
&= -\int (p(z) - \hat{p}_{S\cup S^*}(z))^2 - (p(z) - \hat{p}_S(z))^2 dz\\
&= -\int \left(p(z) - \hat{p}_S(z) - \frac{n}{|S|+n}\left(\hat{p}_{S^*}(z) - \hat{p}_S(z)\right)\right)^2 - (p(z) - \hat{p}_S(z))^2 dz
\end{aligned}$$

$$= -\int \frac{n^2}{(|S|+n)^2} \left(\hat{p}_{S^*}(z) - \hat{p}_S(z)\right)^2 - \frac{2n}{|S|+n} \left\{ (p(z) - \hat{p}_S(z)) \left(\hat{p}_{S^*}(z) - \hat{p}_S(z)\right) \right\} dz.$$

Furthermore,

$$\begin{aligned}(\hat{p}_{S^*}(z) - \hat{p}_S(z))^2 &= (\hat{p}_{S^*}(z) - p(z) + p(z) - \hat{p}_S(z))^2 \\ &= (\hat{p}_{S^*}(z) - p(z))^2 + (p(z) - \hat{p}_S(z))^2 + 2(\hat{p}_{S^*}(z) - p(z))(p(z) - \hat{p}_S(z)), \end{aligned} \tag{12}$$

and

$$\begin{aligned}(p(z) - \hat{p}_S(z))(\hat{p}_{S^*}(z) - \hat{p}_S(z)) &= (p(z) - \hat{p}_S(z))(\hat{p}_{S^*}(z) - p(z) + p(z) - \hat{p}_S(z)) \\ &= (p(z) - \hat{p}_S(z))(\hat{p}_{S^*}(z) - p(z)) + (p(z) - \hat{p}_S(z))^2. \end{aligned} \tag{13}$$

Equations (12) and (13) give

$$\begin{aligned}\mathbb{E}[U(S \cup S^*) - U(S)] = &-\frac{n^2}{(|S|+n)^2} \int (\hat{p}_{S^*}(z) - p(z))^2 dz \\ &+ \frac{n^2 + 2n|S|}{(|S|+n)^2} \int \mathbb{E}[(p(z) - \hat{p}_S(z))^2] dz \\ &+ \frac{2n|S|}{(|S|+n)^2} \int (\hat{p}_{S^*}(z) - p(z)) \mathbb{E}[p(z) - \hat{p}_S(z)] dz. \end{aligned}$$

We can decompose $\mathbb{E}[U(S \cup S^*) - U(S)]$ into two terms by dependency of S^*. To be more specific, $\mathbb{E}[U(S \cup S^*) - U(S)] = h_1(S^*, |S|) + h_2(|S|^*, |S|)$ where

$$h_1(S^*, |S|) = -\frac{n^2}{(|S|+n)^2} \int (\hat{p}_{S^*}(z) - p(z))^2 dz + \frac{2n|S|}{(|S|+n)^2} \int \hat{p}_{S^*}(z) \mathbb{E}[p(z) - \hat{p}_S(z)] dz.$$

Also,

$$\begin{aligned}h_2(n, |S|) &= \frac{n^2 + 2n|S|}{(|S|+n)^2} \int \mathbb{E}[(p(z) - \hat{p}_S(z))^2] dz - \frac{2n|S|}{(|S|+n)^2} \int p(z) \mathbb{E}[p(z) - \hat{p}_S(z)] dz \\ &= \frac{n^2 + 2n|S|}{(|S|+n)^2} \int \mathbb{E}[(p(z) - \hat{p}_S(z))^2] dz - \frac{2n|S|}{(|S|+n)^2} \int p(z)(p(z) - \mathbb{E}[k(z, Z)]) dz. \end{aligned}$$

Therefore, by Ghorbani et al. (2020, Theorem 2.3), we have

$$\begin{aligned}\nu(S^*; U, P, m) &= \frac{1}{m} \sum_{j=1}^{m} \mathbb{E}_{S \sim P^{j-1}}[U(S \cup S^*) - U(S)] \\ &= -\frac{1}{m} \sum_{j=1}^{m} \frac{n^2}{(j+n-1)^2} \int (\hat{p}_{S^*}(z) - p(z))^2 dz \\ &+ \frac{1}{m} \sum_{j=2}^{m} \frac{2n(j-1)}{(j+n-1)^2} \int \hat{p}_{S^*}(z)(p(z) - \mathbb{E}[k(z, Z)]) dz + C_0(n, m) \\ &= -A(n, m) \int (\hat{p}_{S^*}(z) - p(z))^2 dz + B(n, m) g(S^*) + C_0(n, m), \end{aligned} \tag{14}$$

and

$$C_0(n, m) = \frac{1}{m} C_{\text{den}} + \frac{1}{m} \sum_{j=2}^{m} h_2(n, j-1). \tag{15}$$

Hence, it concludes a proof by choosing the constant C_{den} as follows.

$$C_{\text{den}} = -\sum_{j=2}^{m} h_2(n, j-1). \tag{16}$$

□

2 Details for Examples in Section 4

2.1 Details for Example 1

Proof of Example 1. A key idea is to develop Equation (14).

[Step 1] In this step we compute

$$h_2(n,|S|) = \frac{n^2+2n|S|}{(|S|+n)^2}\int \mathbb{E}[(p(z)-\hat{p}_S(z))^2]dz - \frac{2n|S|}{(|S|+n)^2}\int p(z)(p(z)-\mathbb{E}[k(z,Z)])dz.$$

We first compute the term $\int \mathbb{E}[(p(z)-\hat{p}_S(z))^2]dz$. Note that $\hat{p}_S^2(z) = \frac{1}{|S|^2}(\sum_{z_i\in S} k(z,z_i)^2 + \sum_{i\neq j: z_i,z_j\in S} k(z,z_i)k(z,z_j))$. We have

$$\mathbb{E}[\hat{p}_S(z)] = \mathbb{E}[k(z,Z)] = \begin{cases} \frac{1}{2}+\frac{z}{h} & 0\le z\le h/2, \\ 1 & h/2\le z\le 1-h/2, \\ \frac{1}{2}+\frac{1-z}{h} & 1-h/2\le z\le 1, \end{cases}$$

and due to $p(z)=1$,

$$p(z)-\mathbb{E}[k(z,Z)] = \begin{cases} \frac{1}{2}-\frac{z}{h} & 0\le z\le h/2, \\ 0 & h/2\le z\le 1-h/2, \\ \frac{1}{2}-\frac{1-z}{h} & 1-h/2\le z\le 1. \end{cases}$$

Since S are randomly sampled, we have

$$\mathbb{E}[\hat{p}_S^2(z)] = \frac{|S|\mathbb{E}[k(z,Z)]/h + |S|(|S|-1)\mathbb{E}[k(z,Z)]^2}{|S|^2} = \frac{\mathbb{E}[k(z,Z)]}{|S|h} + \frac{|S|-1}{|S|}\mathbb{E}[k(z,Z)]^2.$$

Furthermore, we have $\int \mathbb{E}[k(z,Z)]dz = 1-h/4$ and $\int \mathbb{E}[k(z,Z)]^2dz = 1-5h/12$. Hence, $\int \mathbb{E}[\hat{p}_S^2(z)]dz = \frac{1}{|S|h} - \frac{1}{4|S|} + \frac{|S|-1}{|S|}(1-\frac{5h}{12})$ and we have

$$\begin{aligned}\int \mathbb{E}[(p(z)-\hat{p}_S(z))^2]dz &= 1+\left(\frac{1}{|S|h}-\frac{1}{4|S|}+\frac{|S|-1}{|S|}(1-\frac{5h}{12})\right)-2\left(1-\frac{h}{4}\right) \\ &= \frac{1}{|S|h}-\frac{5}{4|S|}+\frac{(5+|S|)h}{12|S|} \\ &= \frac{12-15h+(5+|S|)h^2}{12|S|h}.\end{aligned}$$

Lastly, $\int p(z)(p(z)-\mathbb{E}[k(z,Z)])dz = h/4$ gives

$$h_2(n,|S|) = \frac{n^2+2n|S|}{(|S|+n)^2}\frac{12-15h+(5+|S|)h^2}{12|S|h} - \frac{2n|S|}{(|S|+n)^2}\frac{h}{4}.$$

[Step 2] By construction of h, $g(S^*)=0$. If $\Delta\ge h$, since z_1^* and z_2^* are apart at least h,

$$\begin{aligned}-\int (p(z)-\hat{p}_{S^*}(z))^2dz &= -\int (1-\hat{p}_{S^*}(z))^2dz \\ &= -\left(|S^*|h\left(1-\frac{1}{|S^*|h}\right)^2+(1-|S^*|h)\right) \\ &= 1-\frac{1}{|S^*|h}.\end{aligned}$$

Therefore, by aggregating all the results in [Step 1] and [Step 2], we have

$$\nu(S^*;U,P,m) = A(2,m)\left(1-\frac{1}{2h}\right)+C_0(2,m).$$

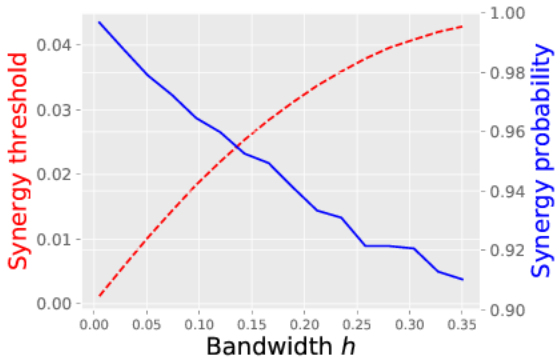

Figure 3: The synergy threshold (red dashed) and the corresponding synergy probability (blue solid) as a function of bandwidth.

Note that by Equation (15),

$$C_0(2,m) = \frac{1}{m}C_{\text{den}} + \frac{1}{m}\sum_{j=2}^{m} h_2(2, j-1).$$

Note that we set $C_{\text{set}} = C_0(2,m)$ in the manuscript.

[Step 3] We now consider the case $\Delta < h$. To this end, without loss of generality, we assume that $z_1^* \le z_2^*$. Then there is overlap between $(z_1^* - h/2, z_1^* + h/2)$ and $(z_2^* - h/2, z_2^* + h/2)$.

$$\hat{p}_{S^*}(z) = \begin{cases} \frac{1}{2h} & z_1^* - h/2 \le z \le z_2^* - h/2, \\ \frac{1}{h} & z_2^* - h/2 \le z \le z_1^* + h/2, \\ \frac{1}{2h} & z_1^* + h/2 \le z \le z_2^* + h/2, \\ 0 & \text{otherwise.} \end{cases}$$

Therefore, $\int (p(z) - \hat{p}_{S^*}(z))^2 dz = -1 + \frac{1}{h} - \frac{\Delta}{2h^2}$. Hence, we have

$$\nu(S^*; U, P, m) = A(2,m)\left(1 - \frac{1}{h} + \frac{\Delta}{2h^2}\right) + C_0(2,m).$$

□

2.2 Details for Example 2

A similar analysis used in Example 1 gives

$$\nu(z_1^*; U, P_Z, m) = A(1,m)(1 - \frac{1}{h}) + C_0(1,m),$$

where $C_0(1,m) = \frac{1}{m}C_{\text{den}} + \frac{1}{m}\sum_{j=2}^{m} h_2(1, j-1)$ by Equation (15). Since it is difficult to solve (8) analytically, we numerically examine when (8) holds when $C_{\text{den}} = 0.2$ and $m = 100$. For fixed bandwidth h, we randomly draw S^* 5000 times and observe if there is a synergy. We empirically find that the synergy is determined by Δ, so we define the synergy threshold as the smallest Δ when the synergy happens, *i.e*, if Δ is greater than the synergy threshold, the inequality $\nu(\{z_1^*, z_2^*\}; U, P_Z, m) \ge \nu(z_1^*; U, P_Z, m) + \nu(z_2^*; U, P_Z, m)$ holds. Also, among the 5000 random sampled sets S^*, we estimate probability that the synergy happens. Figure 3 shows that the synergy threshold and the corresponding synergy probability as a function of h. As h increases, the synergy threshold (in red dashed) increases and the synergy probability (in blue solid) decreases, meaning that in all bandwidths $h \in (0, 0.35)$, the synergy happens when the two points in S^* is far apart to some extent.

3 Implementation details

In this section, we provide implementation details including comprehensive information for algorithms, datasets, and experiment settings. Our implementation codes are available at `https://github.com/ykwon0407/fast_dist_shapley`.

3.1 The proposed algorithms

In order to estimate DShapley, we implicitly assume that we have a set of random samples $\{(\tilde{x}_i, \tilde{y}_i)\}_{i=1}^N$ (*resp.* $\{\tilde{z}_1, \ldots, \tilde{z}_N\}$) from the data distribution $P_{X,Y}$ (*resp.* P_Z). This set is used to estimate unknown quantities. For example, the covariance matrix of inputs Σ_X^{-1} and the squared error e^{*2} for Alg. 1 and Alg. 2, and the optimal bandwidth in kernel for density estimation problem.

Linear regression DShapley in Theorem 2 can be viewed as a cumulative sum of decreasing elements, so the computation of every element would be computationally inefficient. Instead of computing the cumulative sum, we consider the partial sum by ignoring negligible expectation terms. We present a detailed version of Alg. 1.

Algorithm 3 (Detailed) DShapley for the least squares estimator under Gaussian inputs

Require: True value or estimates for $x^{*T}\Sigma_X^{-1}x^*$, e^{*2}, and σ^2. Thresholds $\rho_1 = 0.01, \rho_2 = 0.005$. The maximum number of Monte Carlo samples $T = 10000$. A constant $q \geq p + 3$.

procedure

  Initialize $\hat{\nu}^{\text{old}} \leftarrow 0$

  for $j \in \{q, \ldots, m\}$ **do**

    Initialize $A_j^{\text{old}} \leftarrow 0$

    for $i \in \{1, \ldots, T\}$ **do**

      Sample $t_{[i]}$ from the χ^2_{j-p+1}.

      $A_j^{\text{new}} \leftarrow \left((i-1)A_j^{\text{old}} + \frac{j-1}{j-p}\frac{x^{*T}\Sigma_X^{-1}x^* e^{*2} + t_{[i]}\sigma^2}{(x^{*T}\Sigma_X^{-1}x^* + t_{[i]})^2}\right)/i$ ▷ Based on Theorem 2

      if $|A_j^{\text{new}}/A_j^{\text{old}} - 1| \leq \rho_1$ **then**

        break

      end if

      $A_j^{\text{old}} \leftarrow A_j^{\text{new}}$

    end for

    $\hat{\nu}^{\text{new}} \leftarrow \hat{\nu}^{\text{old}} - A_j^{\text{new}}/m$

    if $|\hat{\nu}^{\text{old}}/\hat{\nu}^{\text{new}} - 1| \leq \rho_2$ **then**

      break

    end if

    $\hat{\nu}^{\text{old}} \leftarrow \hat{\nu}^{\text{new}}$

  end for

  $\hat{\nu}((x^*, y^*); U_q, P_{X,Y}, m) \leftarrow \hat{\nu}^{\text{new}}$ ▷ Estimates for DShapley

end procedure

Binary classification Likewise Alg. 3, the lower bound in Corollary 4 can be viewed as a cumulative sum of decreasing elements, so we again consider the partial sum. A detailed version of Alg. 2 is presented in Alg. 4.

Non-parametric density estimation DShapley in Theorem 5 consists of the two integral terms, $\int(p(z) - \hat{p}_{S^*,k}(z))^2 dz$ and $g(S^*)$. Our approach is to use the MC approximation and to estimate the integrals. Since the first term includes a constant term $\int\{p(z)\}^2 dz$, we ignore the term as in Ghosh (2018, Equation (1.183)). We present a practical example of estimation in Alg. 5.

Accuracy of the proposed algorithms Our algorithms use the Monte-Carlo (MC) method to provide unbiased approximation. When this MC converges, then it guarantees to converge to the true value. In our experiments, we stop the MC when the new increment is small enough compared to the current DShapley estimate

Algorithm 4 (Detailed) DShapley for binary classification

Require: A datum to be valued (x^*, y^*). A set of random samples $\{(X_i, Y_i)\}_{i=1}^B$ from $P_{X,Y}$.
  procedure TRANSFORM_DATA
    while until a convergent condition is met **do**
      $\pi_i \leftarrow \text{logit}^{-1}(X_i^T \hat{\beta}_{\text{IRLS}})$
      Update w_i and Z_i based on Equation (5) and set $\mathbb{W}$ and $\mathbb{Z}$
      $\hat{\beta}_{\text{IRLS}} \leftarrow (\mathbb{X}^T \mathbb{W} \mathbb{X})^{-1} \mathbb{X}^T \mathbb{W} \mathbb{Z}$
    end while
    $\pi^* \leftarrow \text{logit}^{-1}(x^{*T} \hat{\beta}_{\text{IRLS}})$
    $z^* \leftarrow x^{*T} \hat{\beta}_{\text{IRLS}} + (y^* - \pi^*)/(\pi^*(1-\pi^*))$
    $w^* \leftarrow \pi^*(1-\pi^*)$
    Compute a lower bound of DShapley of $\left((w^*)^{1/2} x^*, (w^*)^{1/2} z^*\right)$.
  end procedure
Require: A datum to be valued $((w^*)^{1/2} x^*, (w^*)^{1/2} z^*)$. A set of random samples $\{((w_i)^{1/2} X_i, (w_i)^{1/2} Z_i)\}_{i=1}^B$.
  Hyperparameters $c = C = 1$ and $\rho = 0.005$.
  procedure COMPUTE_LOWER_BOUND
    Initialize $\hat{\nu}^{\text{old}} \leftarrow 0$ and estimate $\tilde{\Sigma}_X$ with $\{(w_1)^{1/2} X_1, \ldots, (w_B)^{1/2} X_B\}$
    $e_{\text{b}}^{*2} \leftarrow (w^*)^{-1}(y^* - \pi^*)^2$
    for $j \in \{q-1, \ldots, m-1\}$ **do**
      $\delta_j \leftarrow (C\sqrt{p} + \sqrt{\frac{\log(jm)}{2c}})/\sqrt{j}$
      $\tilde{\Lambda}_{\text{upper}}(j), \tilde{\Lambda}_{\text{lower}}(j) \leftarrow (j(1-\delta_j)^2)^{-1}, (j(1+\delta_j)^2)^{-1}$
      $\tilde{\Lambda}_{\text{ratio}}(j) \leftarrow \left(\frac{1+w^* x^{*T} \tilde{\Sigma}_X^{-1} x^* \tilde{\Lambda}_{\text{lower}}(j)}{1+w^* x^{*T} \tilde{\Sigma}_X^{-1} x^* \tilde{\Lambda}_{\text{upper}}(j)}\right)^2$
      $A_j \leftarrow \frac{w^* x^{*T} \tilde{\Sigma}_X^{-1} x^* \tilde{\Lambda}_{\text{lower}}^2(j)}{(1+w^* x^{*T} \tilde{\Sigma}_X^{-1} x^* \tilde{\Lambda}_{\text{upper}}(j))^2} \left((2 + w^* x^{*T} \tilde{\Sigma}_X^{-1} x^* \tilde{\Lambda}_{\text{lower}}(j)) - \tilde{\Lambda}_{\text{ratio}}^{-1}(j) e_{\text{b}}^{*2}\right)$ ▷ Based on Corollary 7
      $\hat{\nu}^{\text{new}} \leftarrow \hat{\nu}^{\text{old}} + A_j^{\text{new}}/m$
      if $|\hat{\nu}^{\text{old}}/\hat{\nu}^{\text{new}} - 1| \leq \rho$ **then**
        break
      end if
      $\hat{\nu}^{\text{old}} \leftarrow \hat{\nu}^{\text{new}}$
    end for
  end procedure

to ensure good convergence to the true values. For example, we stop iterations when the new increment is within 0.5% of the current estimates in Alg. 3 or we use large samples ($B = 2000$) in Alg. 5.

3.2 Datasets

Datasets used in time comparison experiment We use the two synthetic datasets for the time comparison experiment (Figure 1) as follows.

- Linear regression: Given (m, p) and $\beta \sim \mathcal{N}(0, I_p)$, we generate $y_i = x_i^T \beta + \epsilon_i$ for all $i \in [m]$. Here, $x_i \sim \mathcal{N}(0, I_p)$ and $\epsilon_i \sim \mathcal{N}(0, 1)$ for all $i \in [m]$. We call this data distribution `Gaussian-R`.
- Binary classification: Given (m, p), we generate $y_i = \textbf{Bern}(0.5)$ and $x_i \sim \mathcal{N}([2 \times y_i, 0, \ldots, 0]^T, I_p)$ for all $i \in [m]$.

Datasets used in point addition experiment We use the two synthetic datasets and eight real datasets for the point addition experiment in Sec. 5. For the synthetic datasets, we generate the two types of datasets, `Gaussian-R` and `Gaussian-C` for regression and classification, respectively. `Gaussian-R` is described above. As for the `Gaussian-C`, we first fix $p = 3$ and set $\beta = (2, 0, 0)$. Then, we generate $x_i \sim \mathcal{N}(0, I_p)$ and $y_i = \textbf{Bern}(\pi_i)$ for all $i \in [m]$. Here $\pi_i := \exp(x_i^T \beta)/(1 + \exp(x_i^T \beta))$. For the real datasets, we collect datasets from multiple sources. For instance, `abalone`, `airfoil`, and `whitewine` are from `UCI Machine Learning Repository` (Dua

Algorithm 5 DShapley for non-parametric density estimation

Require: A set to be valued S^*. A Gaussian kernel k_h. $B = 2000$. A given bandwidth grid $\mathcal{H} := \{h_1, \ldots, h_G\}$. A set of random samples $\{\tilde{z}_1, \ldots, \tilde{z}_B\}$ from P_Z.

procedure

Find optimal bandwidth $h^* \in \mathcal{H}$ which minimizes the five-fold cross-validation error

Set $k \leftarrow k_{h^*}$

Sample $\{\tilde{z}_1^*, \ldots, \tilde{z}_B^*\}$ from $\hat{p}_{S^*,k}$

$\hat{\nu}(S^*; U_k, P_Z, m) \leftarrow -\frac{A(|S^*|,m)}{B} \sum_{i=1}^{B} \left(\hat{p}_{S^*,k}(\tilde{z}_i^*) - 2\hat{p}_{S^*,k}(\tilde{z}_i)\right) + \frac{B(|S^*|,m)}{B} \sum_{i=1}^{B} \left(\hat{p}_{S^*,k}(\tilde{z}_i) - k(\tilde{z}_i^* - \tilde{z}_i)\right)$

end procedure

and Graff, 2017) and `diabetes` is from Efron et al. (2004). A comprehensive list of datasets and details on sample size are provided in Table 3.

For the image datasets `Fashion-MNIST`, `MNIST` and `CIFAR10`, we follow the common procedure in prior works (Ghorbani et al., 2020; Koh and Liang, 2017): we first extract the penultimate layer outputs from the ResNet18 (He et al., 2016) pre-trained with the ImageNet dataset (Russakovsky et al., 2015). After the extraction, we fit the principal component analysis model and extract the first 32 principal components.

Table 3: A summary of datasets for point addition experiment.

Dataset	# of random samples	# of held-out test data	Input dimension	ML problem	Source
`Gaussian-R`	49000	1000	10	Regression	Synthetic dataset
`abalone`	3177	1000	10	Regression	UCI Repository
`airfoil`	1003	500	5	Regression	UCI Repository
`whitewine`	3898	1000	11	Regression	UCI Repository
`Gaussian-C`	49000	1000	3	Classification	Synthetic dataset
`skin-nonskin`	244057	1000	3	Classification	Chang and Lin (2011)
`MNIST`	60000	5000	32	Classification	LeCun et al. (2010)
`diabetes`	342	100	10	Density estimation	Efron et al. (2004)
`australian`	349	100	12	Density estimation	Chang and Lin (2011)
`Fashion-MNIST`	60000	5000	32	Density estimation	Xiao et al. (2017)
`CIFAR10`	50000	5000	32	Classification Density estimation	Krizhevsky et al. (2009)

3.3 Experiment settings

Point addition experiment As for the point addition experiment, we use datasets summarized in Table 3. Throughout the experiments, for each dataset, we first randomly select 200 data points to be valued from datasets. For regression and classification problems, all other data points are used to estimate the DShapley, but for the density estimation problem, we randomly pick 2000 samples. Please note that all the proposed methods, namely Alg. 3, Alg. 4, and Alg. 5, require some data points to estimate unknown-quantities (Σ_X^{-1}, e^{*2}, or bandwidth) Every time point we add a data point given order, we evaluate the test accuracy using the held-out dataset. The held-out dataset sizes are provided in Table 3.

For linear regression cases, we use the utility function constant $C_{\text{lin}} = 2\hat{\sigma}^2$, where $\hat{\sigma} := \frac{1}{m-p} \sum_{i=1}^{m-p} (y_i - x_i^T \hat{\beta})^2$ and $\hat{\beta}$ is the least squares estimator. For classification, the utility function is classification accuracy. Lastly, for density estimation, we considered

$$C_{\text{den}} = mA(n,m) \int \{p(z)\}^2 dz - \sum_{j=2}^{m} h_2(n, j-1),$$

which corresponds to the sum of $mA(n,m) \int \{p(z)\}^2 dz$ and (16), in order to avoid computing $\int \{p(z)\}^2 dz$. As for finding the optimal bandwidth, we select one from $\{10^{-2}, 10^{-1.5}, 10^{-1}, 10^{-0.5}, 10^{0}, 10^{0.5}, 10^{1}\}$ using the five-fold cross-validation error.

4 Additional numerical experiments

4.1 Illustration of DShapley

To see how DShapley changes with respect to $x^{*T}\Sigma_X^{-1}x^*$ and e^{*2}, we estimate DShapley using Algorithm 3. We consider $m \in \{100, 300, 500\}$, $e^{*2} \in \{0, 1, 2, 4, 8\}$, the Gaussian input distribution $X \sim \mathcal{N}_p(0, I_p)$ with $p \in \{10, 30\}$. Here, we assume that Σ_X^{-1} and e^{*2} are known. Figure 4 illustrates DShapley as a function of $x^{*T}\Sigma_X^{-1}x^*$. As anticipated, for a fixed $x^{*T}\Sigma_X^{-1}x^*$, DShapley decreases as e^{*2} increases. Moreover, DShapley exhibits different behavior depending on the error level. When e^{*2} is small, DShapley increases as $x^{*T}\Sigma_X^{-1}x^*$ increases. However, when e^{*2} is big enough, DShapley shows non-monotonic curves in $x^{*T}\Sigma_X^{-1}x^*$. This is because of its form (4). The fraction in (4) has a form of a weighted sum of e^{*2} and σ^2, so it mainly relies on e^{*2} for small values of $x^{*T}\Sigma_X^{-1}x^*$. Lastly, the absolute magnitude of DShapley gets smaller as m increases.

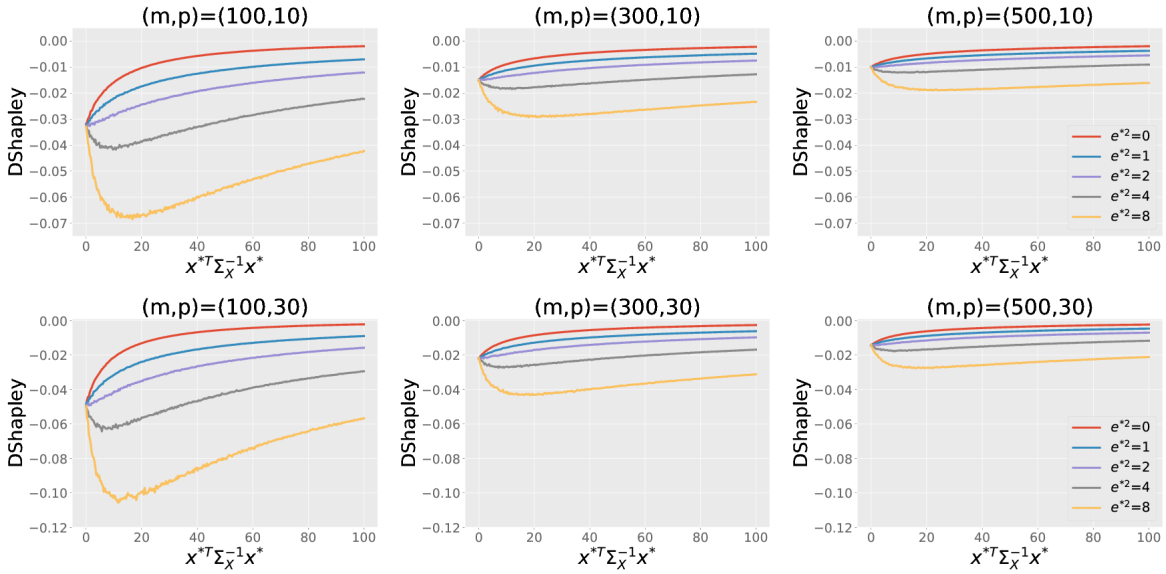

Figure 4: Illustration of DShapley as a function of the Mahalanobis distance $x^{*T}\Sigma_X^{-1}x^*$ when the input dimension p is either (top) 10 or (bottom) 30. Different colors indicate different error levels.

4.2 Point addition experiment with the upper and lower bounds of DShapley

We additionally conduct the point addition experiment with the upper and lower bounds in 3. Although the specific algorithm is not presented, it is straightforward from the 'COMPUTE_LOWER_BOUND' procedure in Alg. 4. We use the same constants C, c, and ρ defined in Alg. 4, but we here set $\gamma = 1/200$.

Figure 5 and Figure 6 show the upper and lower bounds of DShapley when ML problems are regression and classification, respectively. Note that $\mathcal{D}$-SHAPLEY shows the same plots. In our experiments, although the upper bound curves tend to show poor performance, the lower bound curves show promising results. The approximation of DShapley provides computationally efficient solutions, yet this phenomena shows one should be careful when using the approximation based on Theorem 3.

5 A review of Shapley value and its uniqueness

We briefly review the Shapley axioms: symmetry, null player, and additivity. Under the axioms, we describe a fair valuation function (Shapley, 1953). Let U be a utility function and B be a dataset. The three Shapley axioms are symmetry, null player, and additivity defined as follows.

- *Symmetry*: Let $z_i, z_j \in B$. For all $S \subseteq B \backslash \{z_i, z_j\}$, if $U(S \cup \{z_i\}) = U(S \cup \{z_j\})$, then

$$\phi(z_i; U, B) = \phi(z_i; U, B).$$

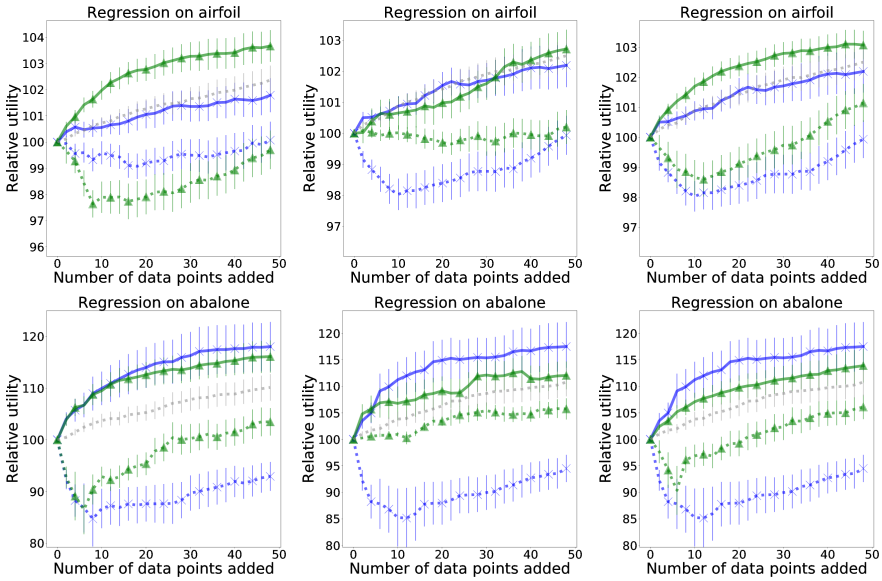

Figure 5: Relative utility and its standard error bar (in %) as a function of the number of data added in linear regression settings. We examine the state-of-the-art $\mathcal{D}$-`SHAPLEY` (blue), random order (gray), and our proposed algorithms (green). As for the proposed algorithms, the exact DShapley based on Theorem 2 (left), the upper (center), and the lower bounds based on Theorem 3 (right). The solid and dashed curves correspond to adding points with the largest and smallest values first, respectively. The results are based on 50 repetitions.

- *Null player*: Let $z_i \in B$. For all $S \subseteq B \backslash \{z_i\}$, if $U(S \cup \{z_i\}) = U(S)$, then

$$\phi(z_i; U, B) = 0.$$

- *Additivity*: Let U_1, U_2 be two utility functions. For all $z \in B$,

$$\phi(z; U_1 + U_2, B) = \phi(z; U_1, B) + \phi(z; U_2, B).$$

Under the axioms, we provide the following uniqueness theorem quote from Osborne and Rubinstein (1994, Proposition 293.1).

Theorem 8. *Under the three Shapley axioms, the Shapley value is the unique valuation.*

References

Chang, C.-C. and Lin, C.-J. (2011). Libsvm: A library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, 2(3):1–27.

Dua, D. and Graff, C. (2017). UCI machine learning repository.

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., et al. (2004). Least angle regression. *The Annals of statistics*, 32(2):407–499.

Ghorbani, A., Kim, M. P., and Zou, J. (2020). A distributional framework for data valuation. *arXiv preprint arXiv:2002.12334*.

Ghosh, S. (2018). *Kernel smoothing: Principles, methods and applications*. John Wiley & Sons.

Gupta, A. K. and Nagar, D. K. (1999). *Matrix variate distributions*, volume 104. CRC Press.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

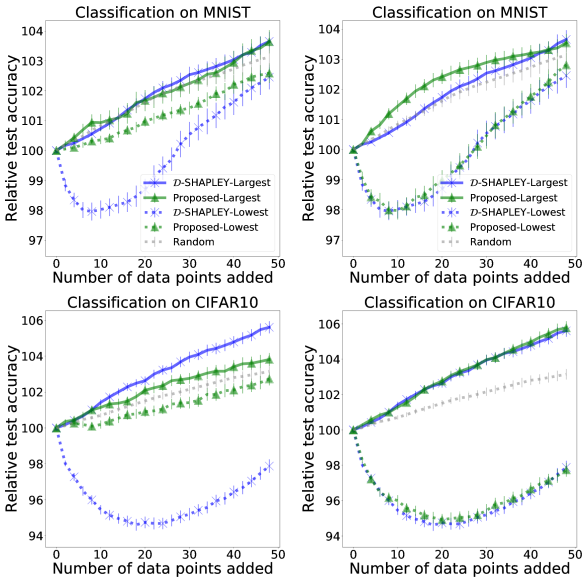

Figure 6: Relative utility and its standard error bar (in %) as a function of the number of data added in classification settings. We examine the state-of-the-art $\mathcal{D}$-SHAPLEY (blue), random order (gray), and our proposed algorithms (green). As for the proposed algorithms, the upper (left) and lower bounds based on Corollary 7 (right). The solid and dashed curves correspond to adding points with the largest and smallest values first, respectively. The results are based on 50 repetitions.

Koh, P. W. and Liang, P. (2017). Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1885–1894. JMLR. org.

Krizhevsky, A. et al. (2009). Learning multiple layers of features from tiny images.

LeCun, Y., Cortes, C., and Burges, C. (2010). Mnist handwritten digit database. *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, 2.

Osborne, M. J. and Rubinstein, A. (1994). *A course in game theory*.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. (2015). Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252.

Shapley, L. S. (1953). A value for n-person games. *Contributions to the Theory of Games*, 2(28):307–317.

Vershynin, R. (2010). Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*.

Xiao, H., Rasul, K., and Vollgraf, R. (2017). Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*.