RDM-DC: Poisoning Resilient Dataset Condensation with Robust Distribution Matching (Supplementary Material)

Tianhang Zheng[1] **Baochun Li**[1]

[1]Department of Electrical and Computer Engineering, University of Toronto

1 OMITTED PROOF

Lemma 1.1 *Assuming that* $\mathcal{D}$ *and* $\mathcal{B}$ *have bounded covariance matrices* $\boldsymbol{\Sigma}_{\mathcal{D}}, \boldsymbol{\Sigma}_{\mathcal{B}} \leq \sigma^2 \boldsymbol{I}$*, and their means have an apparent difference,* i.e., $\|\boldsymbol{\mu}_{\mathcal{D}} - \boldsymbol{\mu}_{\mathcal{B}}\|_2^2 \geq \frac{\alpha\sigma^2}{\epsilon}$ *where* $\alpha > \frac{2665}{576}$*, then if we drop all the representations that satisfies* $|\langle \boldsymbol{r} - \boldsymbol{\mu}_{\mathcal{P}}, \boldsymbol{v}\rangle| \geq t$ *with a certain* t*, then we can reduce the scale of the poisoned deviation from* $O(\epsilon\sqrt{d_{\boldsymbol{r}}})$ *to* $\Theta(\epsilon^2\sqrt{d_{\boldsymbol{r}}})$.

To prove the above lemma, we need the help of Chebyshev's inequality, which is introduced in the following.

Lemma 1.2 (Chebyshev's inequality) *Given a scalar random variable* X*, if* $\mathbb{E}[X] = \mu$ *and* $Var[X] = \sigma^2$*, then*

$$\mathbb{P}(|X - \mu| \geq t) \leq \frac{\sigma^2}{t^2} \tag{1}$$

Given Chebyshev's inequality, we have the following corollary, which will be used in the proof of Lemma 1.1.

Corollary 1.1 *Given a multi-dimensional variable* $\boldsymbol{X}$*, if* $\mathbb{E}[\boldsymbol{X}] = \boldsymbol{\mu}$ *and* $Cov[\boldsymbol{X}] \leq \sigma^2 \boldsymbol{I}$*, then for any unit vector* $\boldsymbol{u}$*, we have*

$$\mathbb{P}(|\langle \boldsymbol{X} - \boldsymbol{\mu}, \boldsymbol{u}\rangle| > t) \leq \frac{\sigma^2}{t^2} \tag{2}$$

Proof [Proof of Corollary 1.1]

Considering $\langle \boldsymbol{X}, \boldsymbol{u}\rangle$ as a scalar random variable, we have $\mathbb{E}[\langle \boldsymbol{X}, \boldsymbol{u}\rangle] = \langle \boldsymbol{\mu}, \boldsymbol{u}\rangle$ and,

$$\text{Var}[\langle \boldsymbol{X}, \boldsymbol{u}\rangle] = \boldsymbol{u}^T Cov[\boldsymbol{X}]\boldsymbol{u} \leq \sigma^2. \tag{3}$$

With Chebyshev's inequality, we know that

$$\mathbb{P}(|\langle \boldsymbol{X}, \boldsymbol{u}\rangle - \langle \boldsymbol{\mu}, \boldsymbol{u}\rangle| \geq t) \leq \frac{\text{Var}[\langle \boldsymbol{X}, \boldsymbol{u}\rangle]}{t^2} \leq \frac{\sigma^2}{t^2} \tag{4}$$

■

Beyond Corollary 1.1, we also need to use the following lemma and corollary in the proof of Lemma 1.1.

Lemma 1.3 *Given two distributions* $\boldsymbol{P}$ *and* $\boldsymbol{Q}$ *with mean* $\boldsymbol{\mu_P}$ *and* $\boldsymbol{\mu_Q}$ *and covariance matrices* $\boldsymbol{\Sigma_P}, \boldsymbol{\Sigma_Q} \leq \sigma^2 \boldsymbol{I}$*, if* $\|\boldsymbol{\mu_P} - \boldsymbol{\mu_Q}\|_2^2 \geq \frac{\alpha\sigma^2}{\epsilon}$*, then* $\langle \boldsymbol{v}, \boldsymbol{\mu_P} - \boldsymbol{\mu_Q}\rangle^2 \geq \frac{\alpha\sigma^2 - \sigma^2/(1-\epsilon)}{\epsilon}$ *where* $\boldsymbol{v}$ *is the first eigenvector of the covariance matrix of* $(1-\epsilon)\boldsymbol{P} + \epsilon\boldsymbol{Q}$.

Proof [Proof of Lemma 1.3] The mean of the mixture $(1-\epsilon)\boldsymbol{P} + \epsilon\boldsymbol{Q}$ is $(1-\epsilon)\boldsymbol{\mu_P} + \epsilon\boldsymbol{\mu_Q}$, which is denoted by $\boldsymbol{\mu_M}$. We denote $\boldsymbol{\mu_P} - \boldsymbol{\mu_Q}$ by $\boldsymbol{\delta}$. The covariance matrix of $(1-\epsilon)\boldsymbol{P} + \epsilon\boldsymbol{Q}$ can be expressed as

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{X}\sim(1-\epsilon)\boldsymbol{P}+\epsilon\boldsymbol{Q}}[(\boldsymbol{X} - \boldsymbol{\mu_M})(\boldsymbol{X} - \boldsymbol{\mu_M})^T] \\
&\qquad = (1-\epsilon)\mathbb{E}_{\boldsymbol{X}\sim\boldsymbol{P}}[(\boldsymbol{X} - \boldsymbol{\mu_M})(\boldsymbol{X} - \boldsymbol{\mu_M})^T] \\
&\qquad\quad + \epsilon\mathbb{E}_{\boldsymbol{X}\sim\boldsymbol{Q}}[(\boldsymbol{X} - \boldsymbol{\mu_M})(\boldsymbol{X} - \boldsymbol{\mu_M})^T]
\end{aligned} \tag{5}$$

Since we have

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{X}\sim\boldsymbol{P}}[(\boldsymbol{X} - \boldsymbol{\mu_M})(\boldsymbol{X} - \boldsymbol{\mu_M})^T] \\
&\quad = \mathbb{E}_{\boldsymbol{X}\sim\boldsymbol{P}}[(\boldsymbol{X} - \boldsymbol{\mu_P} + \epsilon\boldsymbol{\delta})(\boldsymbol{X} - \boldsymbol{\mu_P} + \epsilon\boldsymbol{\delta})^T] \\
&\quad = \boldsymbol{\Sigma_P} + \epsilon^2\boldsymbol{\delta}\boldsymbol{\delta}^T \\
&\mathbb{E}_{\boldsymbol{X}\sim\boldsymbol{Q}}[(\boldsymbol{X} - \boldsymbol{\mu_M})(\boldsymbol{X} - \boldsymbol{\mu_M})^T] \\
&\quad = \mathbb{E}_{\boldsymbol{X}\sim\boldsymbol{Q}}[(\boldsymbol{X} - \boldsymbol{\mu_Q} - (1-\epsilon)\boldsymbol{\delta})^T)(\boldsymbol{X} - \boldsymbol{\mu_Q} - (1-\epsilon)\boldsymbol{\delta})^T] \\
&\quad = \boldsymbol{\Sigma_Q} + (1-\epsilon)^2\boldsymbol{\delta}\boldsymbol{\delta}^T\ ,
\end{aligned}$$

we have a lower bound on the covariance matrix of the mixture $(1-\epsilon)\boldsymbol{P} + \epsilon\boldsymbol{Q}$ as

$$\begin{aligned}
\boldsymbol{\Sigma_M} &= \mathbb{E}_{\boldsymbol{X}\sim(1-\epsilon)\boldsymbol{P}+\epsilon\boldsymbol{Q}}[(\boldsymbol{X} - \boldsymbol{\mu_M})(\boldsymbol{X} - \boldsymbol{\mu_M})^T] \\
&= (1-\epsilon)\boldsymbol{\Sigma_P} + \epsilon\boldsymbol{\Sigma_Q} + \epsilon(1-\epsilon)\boldsymbol{\delta}\boldsymbol{\delta}^T \geq \epsilon(1-\epsilon)\boldsymbol{\delta}\boldsymbol{\delta}^T.
\end{aligned} \tag{6}$$

Accepted for the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023).

Suppose that $\boldsymbol{v}$ is the first eigenvector of $\boldsymbol{\Sigma_M}$ and $\boldsymbol{u} = \frac{\boldsymbol{\delta}}{\|\boldsymbol{\delta}\|_2}$, we then have

$$\boldsymbol{v}^T\boldsymbol{\Sigma_M}\boldsymbol{v} \geq \boldsymbol{u}^T\boldsymbol{\Sigma_M}\boldsymbol{u} \geq \epsilon(1-\epsilon)\boldsymbol{u}^T\boldsymbol{\delta}\boldsymbol{\delta}^T\boldsymbol{u} = \epsilon(1-\epsilon)\|\boldsymbol{\delta}\|_2^2. \tag{7}$$

Since $\boldsymbol{\Sigma_P}, \boldsymbol{\Sigma_Q} \leq \sigma^2\boldsymbol{I}$, we also have

$$\begin{aligned}\boldsymbol{v}^T\boldsymbol{\Sigma_M}\boldsymbol{v} &= (1-\epsilon)\boldsymbol{v}^T\boldsymbol{\Sigma_P}\boldsymbol{v} + \epsilon\boldsymbol{v}^T\boldsymbol{\Sigma_Q}\boldsymbol{v} + \epsilon(1-\epsilon)\boldsymbol{v}^T\boldsymbol{\delta}\boldsymbol{\delta}^T\boldsymbol{v} \\ &\leq \sigma^2 + \epsilon(1-\epsilon)\langle\boldsymbol{v},\boldsymbol{\delta}\rangle^2 \end{aligned} \tag{8}$$

Thus, we have

$$\langle\boldsymbol{v},\boldsymbol{\delta}\rangle^2 \geq \frac{\boldsymbol{v}^T\boldsymbol{\Sigma_M}\boldsymbol{v} - \sigma^2}{\epsilon(1-\epsilon)} \geq \|\boldsymbol{\delta}\|_2^2 - \frac{\sigma^2}{\epsilon(1-\epsilon)} \tag{9}$$

Given the assumption that $\|\boldsymbol{\delta}\|_2^2 \geq \frac{\alpha\sigma^2}{\epsilon}$,

$$\langle\boldsymbol{v},\boldsymbol{\delta}\rangle^2 \geq \frac{\alpha\sigma^2 - \sigma^2/(1-\epsilon)}{\epsilon} \tag{10}$$

■

Based on Lemma 1.3, we have the following corollary.

Corollary 1.2 *Given the definitions and conditions in Lemma 1.3, if* $\epsilon \leq \frac{1}{10}$ *and* $\alpha > \frac{2665}{576}$*, then we have* $(1-2\epsilon)|\langle\boldsymbol{\delta},\boldsymbol{v}\rangle| > \frac{3\sigma}{2\sqrt{\epsilon}}$.

Proof Given Lemma 1.3, we have

$$(1-2\epsilon)|\langle\boldsymbol{\delta},\boldsymbol{v}\rangle| \geq (1-2\epsilon)\sqrt{\alpha - \frac{1}{1-\epsilon}}\frac{\sigma}{\sqrt{\epsilon}} \tag{11}$$

Since $1-2\epsilon$ and $-\frac{1}{1-\epsilon}$ are decreasing functions w.r.t. ϵ, they achieve the minimum at $\epsilon = \frac{1}{10}$. Thus, we have

$$(1-2\epsilon)|\langle\boldsymbol{\delta},\boldsymbol{v}\rangle| \geq \frac{4}{5}\sqrt{\alpha - \frac{10}{9}}\frac{\sigma}{\sqrt{\epsilon}}. \tag{12}$$

So if $\alpha > \frac{2665}{576}$, we have $(1-2\epsilon)|\langle\boldsymbol{\delta},\boldsymbol{v}\rangle| > \frac{3\sigma}{2\sqrt{\epsilon}}$. ■

Proof [Proof of Lemma 1.1] The mean of the poisoned representation distribution $\mathcal{P}$ is $\boldsymbol{\mu}_{\mathcal{P}} = (1-\epsilon)\boldsymbol{\mu}_{\mathcal{D}} + \epsilon\boldsymbol{\mu}_{\mathcal{B}}$. Let $\boldsymbol{\delta} = \boldsymbol{\mu}_{\mathcal{B}} - \boldsymbol{\mu}_{\mathcal{D}}$ and $t = |\epsilon\langle\boldsymbol{\delta},\boldsymbol{v}\rangle| + \frac{\sigma}{\sqrt{\epsilon}}$. We denote the covariance matrix of $\mathcal{P}$ by $\boldsymbol{\Sigma}_{\mathcal{P}}$ and its first eigenvector by $\boldsymbol{v}$.

For the original representation distribution, we have

$$\begin{aligned}&\mathbb{P}_{\boldsymbol{r}\sim\mathcal{D}}[|\langle\boldsymbol{r}-\boldsymbol{\mu}_{\mathcal{P}},\boldsymbol{v}\rangle| > t] \\ &\quad= \mathbb{P}_{\boldsymbol{r}\sim\mathcal{D}}[|\langle\boldsymbol{r}-\boldsymbol{\mu}_{\mathcal{D}},\boldsymbol{v}\rangle - \epsilon\langle\boldsymbol{\delta},\boldsymbol{v}\rangle| > t] \; ① \\ &\quad\leq \mathbb{P}_{\boldsymbol{r}\sim\mathcal{D}}[|\langle\boldsymbol{r}-\boldsymbol{\mu}_{\mathcal{D}},\boldsymbol{v}\rangle| > \frac{\sigma}{\sqrt{\epsilon}}] \; ② \\ &\quad\leq \epsilon \; ③ \end{aligned} \tag{13}$$

① is because $\boldsymbol{\mu}_{\mathcal{P}} = \boldsymbol{\mu}_{\mathcal{D}} + \epsilon\boldsymbol{\delta}$. ② is because if $|\langle\boldsymbol{r}-\boldsymbol{\mu}_{\mathcal{D}},\boldsymbol{v}\rangle - \epsilon\langle\boldsymbol{\delta},\boldsymbol{v}\rangle| > t$, then either $\langle\boldsymbol{r}-\boldsymbol{\mu}_{\mathcal{D}}\rangle > t + \epsilon\langle\boldsymbol{\delta},\boldsymbol{v}\rangle > \frac{\sigma}{\sqrt{\epsilon}}$ or $\langle\boldsymbol{r}-\boldsymbol{\mu}_{\mathcal{D}}\rangle < -t + \epsilon\langle\boldsymbol{\delta},\boldsymbol{v}\rangle < -\frac{\sigma}{\sqrt{\epsilon}}$ holds true. Thus, we have $|\langle\boldsymbol{r}-\boldsymbol{\mu}_{\mathcal{D}}\rangle| > \frac{\sigma}{\sqrt{\epsilon}}$, and $\{\boldsymbol{r}, |\langle\boldsymbol{r}-\boldsymbol{\mu}_{\mathcal{D}},\boldsymbol{v}\rangle - \epsilon\langle\boldsymbol{\delta},\boldsymbol{v}\rangle| > t\} \subseteq \{\boldsymbol{r}, |\langle\boldsymbol{r}-\boldsymbol{\mu}_{\mathcal{D}},\boldsymbol{v}\rangle| > \frac{\sigma}{\sqrt{\epsilon}}\}$. Therefore, ② holds true. ③ is because of Corollary 1.1.

For the poisoned distribution, we have

$$\begin{aligned}&\mathbb{P}_{\boldsymbol{r}\sim\mathcal{B}}[|\langle\boldsymbol{r}-\boldsymbol{\mu}_{\mathcal{P}},\boldsymbol{v}\rangle| < t] \\ &\quad= \mathbb{P}_{\boldsymbol{r}\sim\mathcal{B}}[|\langle\boldsymbol{r}-\boldsymbol{\mu}_{\mathcal{B}},\boldsymbol{v}\rangle + (1-\epsilon)\langle\boldsymbol{\delta},\boldsymbol{v}\rangle| < t] \; ① \\ &\quad\leq \mathbb{P}_{\boldsymbol{r}\sim\mathcal{B}}[|\langle\boldsymbol{r}-\boldsymbol{\mu}_{\mathcal{B}},\boldsymbol{v}\rangle| > (1-2\epsilon)|\langle\boldsymbol{\delta},\boldsymbol{v}\rangle| - \frac{\sigma}{\sqrt{\epsilon}}] \; ② \\ &\quad\leq \mathbb{P}_{\boldsymbol{r}\sim\mathcal{B}}[|\langle\boldsymbol{r}-\boldsymbol{\mu}_{\mathcal{B}},\boldsymbol{v}\rangle| > \frac{\sigma}{2\sqrt{\epsilon}}] \leq 4\epsilon \; ③ \end{aligned} \tag{14}$$

① is because $\boldsymbol{\mu}_{\mathcal{P}} = \boldsymbol{\mu}_{\mathcal{B}} - (1-\epsilon)\boldsymbol{\delta}$. In the following, we prove ②: Given $|\langle\boldsymbol{r}-\boldsymbol{\mu}_{\mathcal{B}},\boldsymbol{v}\rangle + (1-\epsilon)\langle\boldsymbol{\delta},\boldsymbol{v}\rangle| < t$, we have $-t-(1-\epsilon)\langle\boldsymbol{\delta},\boldsymbol{v}\rangle < \langle\boldsymbol{r}-\boldsymbol{\mu}_{\mathcal{B}},\boldsymbol{v}\rangle < t-(1-\epsilon)\langle\boldsymbol{\delta},\boldsymbol{v}\rangle$. Since $t = |\epsilon\langle\boldsymbol{\delta},\boldsymbol{v}\rangle| + \frac{\sigma}{\sqrt{\epsilon}}$, $-|\epsilon\langle\boldsymbol{\delta},\boldsymbol{v}\rangle| - \frac{\sigma}{\sqrt{\epsilon}} - (1-\epsilon)\langle\boldsymbol{\delta},\boldsymbol{v}\rangle < \langle\boldsymbol{r}-\boldsymbol{\mu}_{\mathcal{B}},\boldsymbol{v}\rangle < |\epsilon\langle\boldsymbol{\delta},\boldsymbol{v}\rangle| + \frac{\sigma}{\sqrt{\epsilon}} - (1-\epsilon)\langle\boldsymbol{\delta},\boldsymbol{v}\rangle$.

Then, we consider two cases: If $\langle\boldsymbol{\delta},\boldsymbol{v}\rangle \geq 0$, we have $\langle\boldsymbol{r}-\boldsymbol{\mu}_{\mathcal{B}},\boldsymbol{v}\rangle < \frac{\sigma}{\sqrt{\epsilon}} - (1-2\epsilon)|\langle\boldsymbol{\delta},\boldsymbol{v}\rangle|$. Given Corollary 1.2, we have $|\langle\boldsymbol{r}-\boldsymbol{\mu}_{\mathcal{B}},\boldsymbol{v}\rangle| > (1-2\epsilon)|\langle\boldsymbol{\delta},\boldsymbol{v}\rangle| - \frac{\sigma}{\sqrt{\epsilon}}$. If $\langle\boldsymbol{\delta},\boldsymbol{v}\rangle < 0$, we have $\langle\boldsymbol{r}-\boldsymbol{\mu}_{\mathcal{B}},\boldsymbol{v}\rangle > (1-2\epsilon)|\langle\boldsymbol{\delta},\boldsymbol{v}\rangle| - \frac{\sigma}{\sqrt{\epsilon}}$. Given Corollary 1.2, we also have $|\langle\boldsymbol{r}-\boldsymbol{\mu}_{\mathcal{B}},\boldsymbol{v}\rangle| > (1-2\epsilon)|\langle\boldsymbol{\delta},\boldsymbol{v}\rangle| - \frac{\sigma}{\sqrt{\epsilon}}$. Therefore, ② holds true. ③ is because of Corollary 1.2.

Suppose after filtering out the data that satisfies $|\langle\boldsymbol{r}-\boldsymbol{\mu}_{\mathcal{P}},\boldsymbol{v}\rangle| \geq t$, the remaining deviation caused by $\mathcal{B}$ is expected to be

$$\begin{aligned}&|\epsilon\mathbb{E}_{\boldsymbol{r}\sim\mathcal{B},|\langle\boldsymbol{r}-\boldsymbol{\mu}_{\mathcal{P}},\boldsymbol{v}\rangle|<t}[\boldsymbol{r}]| < \epsilon t\mathbb{P}_{\boldsymbol{r}\sim\mathcal{B}}[|\langle\boldsymbol{r}-\boldsymbol{\mu}_{\mathcal{P}},\boldsymbol{v}\rangle| < t] \\ &\quad\leq 4\epsilon^2 t = 4\epsilon^2(|\epsilon\langle\boldsymbol{\delta},\boldsymbol{v}\rangle| + \frac{\sigma}{\sqrt{\epsilon}}) \end{aligned} \tag{15}$$

Since $\frac{\sigma}{\sqrt{\epsilon}} \leq \frac{2}{3}|\epsilon\langle\boldsymbol{\delta},\boldsymbol{v}\rangle|$ according to Corollary 1.2, we have

$$|\epsilon\mathbb{E}_{\boldsymbol{r}\sim\mathcal{B},|\langle\boldsymbol{r}-\boldsymbol{\mu}_{\mathcal{P}},\boldsymbol{v}\rangle|<t}[\boldsymbol{r}]| \leq \frac{20}{3}\epsilon^3|\langle\boldsymbol{\delta},\boldsymbol{v}\rangle| \leq \frac{20}{3}\epsilon^3\|\boldsymbol{\delta}\|_2 \tag{16}$$

Since $\epsilon \leq \frac{1}{10}$ and $\|\boldsymbol{\delta}\|_2 \sim \Theta(\sqrt{d_{\boldsymbol{r}}})$, we have

$$|\epsilon\mathbb{E}_{\boldsymbol{r}\sim\mathcal{B},|\langle\boldsymbol{r}-\boldsymbol{\mu}_{\mathcal{P}},\boldsymbol{v}\rangle|<t}[\boldsymbol{r}]| \sim \Theta(\epsilon^2\sqrt{d_{\boldsymbol{r}}}). \tag{17}$$

■