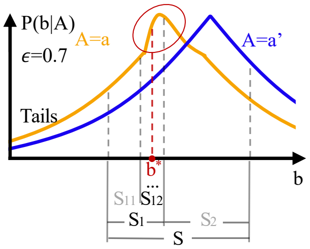


	Pr[M(a)∈S]	Pr[M(a')∈S]	ξ(a,a',S)
S1	0.3	0.2	ξ_1=0.41
S2	0.4	0.3	ξ_2=0.29
S11	0.12	0.1	ξ_{11}=0.18
S12	0.18	0.1	ξ_{12}=0.59

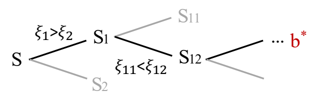


Figure 15: Smaller events can achieve larger distinguishability. $M(a')$ follows a Laplace-shaped distribution, while $M(a)$ shows a tilted density surge marked with the red circle. Specific probability values are listed in the table, and the power ξ is calculated accordingly. The maximum ξ happens at the extreme event b^*, which can be approached by gradually shrinking the event $S \to S_1 \to S_{12} \to \cdots \to b^*$ as in the flow chart.

A SMALLER EVENTS ACHIEVE LARGER DISTINGUISHABILITY

We show that for $\delta_c = 0$, the auditing scheme can always increase its power by progressively shrinking its outcome set S. Specifically, consider an outcome S that can be partitioned into two disjoint subsets S_1 and S_2 as in Figure 15. Then we have $\xi(a,a',S) \le \max\{\xi(a,a',S_1), \xi(a,a',S_2)\}$, with the following proof.

Proof.

$$\begin{aligned}
\xi &= \ln\left(\frac{\Pr[M(a)\in S]}{Pr[M(a')\in S]}\right)\\
&= \ln\left(\frac{\Pr[M(a)\in S_1]+\Pr[M(a)\in S_2]}{Pr[M(a')\in S_1]+Pr[M(a')\in S_2]}\right)\\
&\le \ln\left(\max(\frac{\Pr[M(a)\in S_1]}{\Pr[M(a')\in S_1]}, \frac{\Pr[M(a)\in S_2]}{\Pr[M(a)\in S_2]})\right)\\
&= \max\left(\ln(\frac{\Pr[M(a)\in S_1]}{\Pr[M(a')\in S_1]}), \ln(\frac{\Pr[M(a)\in S_2]}{\Pr[M(a')\in S_2]})\right)\\
&= \max(\xi_1, \xi_2).
\end{aligned}$$

□

B PROOF OF PREREQUISITE 1

Given any mechanism M_ϑ, the iff condition for $\xi^*(\vartheta) < \epsilon^*(\vartheta)$ against DP-Sniper's auditing with probability threshold c is

$$\Pr[M_\vartheta(a') \in S^*] < c, \tag{17}$$

where S^* is the theoretical optimal outcome set in Eq.(3). That is, M_ϑ must satisfy Eq.(7) to produce a false positive against DP-Sniper.

Proof. First, we prove sufficiency. From Eq. (17), we can infer that $S^* \in \hat{S}$. To satisfy $\Pr[M(a') \in \hat{S}] = c$, we let the complementary set of S^* in $\hat{S}$ be S'. According to the definition of S^*, we can obtain that for all $b' \in S', b^* \in S^*, r(b') < r(b^*)$ which is

$$\frac{\Pr[M(a)\in S']}{\Pr[M(a')\in S']} < \frac{\Pr[M(a)\in S^*]}{\Pr[M(a')\in S^*]}.$$

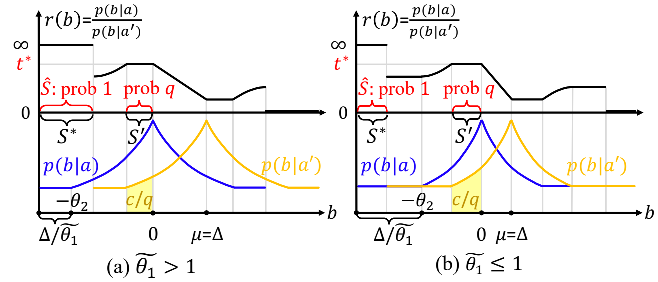


Figure 16: DP-Sniper's $\hat{S}$ against the adapted Laplace $M^{\text{lap}}_{\widetilde{\theta}}$. The theoretical optimal set is $S^* = [-\widetilde{\theta_2} - \Delta/\widetilde{\theta_1}, -\widetilde{\theta_2})$ or $S^* = [-\widetilde{\theta_2} - \Delta/\widetilde{\theta_1}, -\widetilde{\theta_2} - \Delta/\widetilde{\theta_1} + \Delta)$ depending on whether $\widetilde{\theta_1} > 1$.

Based on a simple inequality, we can obtain the following conclusion:

$$\begin{aligned}
e^{\xi^*(\vartheta)} &= \frac{\Pr[M(a)\in S'] + \Pr[M(a)\in S^*]}{\Pr[M(a')\in S'] + \Pr[M(a')\in S^*]}\\
&< \frac{\Pr[M(a)\in S^*]}{\Pr[M(a')\in S^*]} = e^{\epsilon^*(\vartheta)}.
\end{aligned}$$

This directly yields $\xi^*(\vartheta) < \epsilon^*(\vartheta)$.

Next, we proceed to prove necessity. Using proof by contradiction, suppose the conclusion is not true, meaning that $\Pr[M_\vartheta(a') \in S^*] >= c$ holds. Beacuse of $\forall b^* \in S^*, r(b^*) \equiv r^*$, we must have $t = r^*$. Otherwise, it would lead to the non-fulfillment of $\Pr[M(a') \in \hat{S}] = c$. This implies $\hat{S} \in S^*$. Eventually, we can derive $\xi^*(\vartheta) < \epsilon^*(\vartheta) = r^*$, which contradicts the initial conditions. Therefore, $\Pr[M_\vartheta(a') \in S^*] < c$ holds. □

C PROOF FOR SEC. 5.1

The proof of Thm. 2:

Proof. We show how this adapted mechanism $M^{\text{lap}}_{\widetilde{\theta}}$ aligns with P1, R1 and R2. The conditions differ depending on the value of $\widetilde{\theta_1}$.
(P1): As illustrated in Figure 16, the maximal likelihood ratio is $r(b) = \infty$, achieved at S^* where the density of a' equals zero. Consequently, $\Pr[M_{\widetilde{\theta}}(a') \in S^*] = 0$, thereby naturally fulfilling prerequisite 1.
(R1): The theoretical DP level is $\epsilon^*(\widetilde{\theta}) = \xi(a,a',S^*) = \infty$. Hence R1 is also naturally satisfied for any $\widetilde{\theta}$ value.
(R2): We first identify DP-Sniper's $\hat{S}$ against the adapted mechanism. As illustrated in Figure 16, the ratio $r(b)$ loses its monotonicity and the outcome set with the second largest ratio is $S' := [-\widetilde{\theta_2}+\Delta, 0]$. To simplify computation, we only consider $\widetilde{\theta}$ for which $\Pr[M^{\text{lap}}_{\widetilde{\theta}}(a') \in S^* \cup S'] \ge c$, in which case

$$\Pr[b \in \hat{S}] = \mathbb{1}[b \in S^*] + q \cdot \mathbb{1}[b \in S'], \ q = \frac{c}{\Pr[M^{\text{lap}}_{\widetilde{\theta}}(a') \in S']}.$$

Then $\xi^*(\widetilde{\theta}) = \ln(\Pr[M^{\text{lap}}_{\widetilde{\theta}}(a) \in S^*] + q \cdot \Pr[M^{\text{lap}}_{\widetilde{\theta}}(a) \in S']) - \ln(c)$. Specifically, if $\widetilde{\theta_1} > 1$, $S^* = [-\widetilde{\theta_2} - \frac{\Delta}{\widetilde{\theta_1}}, -\widetilde{\theta_2})$ as in Figure 16(a) and R2 becomes Eq.(9a); otherwise, $S^* = [-\widetilde{\theta_2} - \frac{\Delta}{\widetilde{\theta_1}}, -\widetilde{\theta_2} - \frac{\Delta}{\widetilde{\theta_1}} + \Delta)$ as in Figure 16(b), and R2 becomes Eq.(9b). Detailed derivations are omitted for brevity.

Combining the analysis above, any $M^{\text{lap}}_{\widetilde{\theta}}$ that satisfies Eq.(9) is an FP. This is a sufficient but not necessary condition: the empirical $\hat{S}$ can take on other formulations, so there can be other false positive instances besides the one instantiated here. □

D PROOF FOR SEC. 5.2

Before addressing P1 and R1, we notice that the probabilities of b^0 and $S'' := \{b^j | j \neq 0\}$ are relatively easy to compute. This observation leads us to a key simplification:

Lemma 1. *Given $S'' := \{b^j | j \neq 0\}$, we have $r(b^j) > r(b^0)$ for all $j \in [1, N]$. Therefore, $S^* \subset S''$. It further follows that $\Pr[M^{svt}_{\widetilde{\theta}}(a') \in S''] > \Pr[M^{svt}_{\widetilde{\theta}}(a') \in S^*]$ and $\xi(a, a', S'') < \epsilon^*(\widetilde{\theta})$.*

The proof of Lemma 1 is deferred to the end of this subsection. With Lemma 1, we can replace P1 and R1 with simpler surrogates: P1 is relaxed to $\Pr[M^{\text{svt}}_{\widetilde{\theta}}(a') \in S''] < c$, and R1 is relaxed to $\xi(a, a', S'') > \epsilon_c$. This approach eliminates the need to pinpoint the exact S^* or calculate the intricate $\epsilon^*(\widetilde{\theta})$, allowing our analysis to concentrate on S'' or b^0.

(P1): To satisfy the original prerequisite 1, it suffices to ensure $\Pr[M^{\text{svt}}_{\widetilde{\theta}}(a') \in S''] < c$, which is specified as Eq.(11a).

(R1): We solve its surrogate of $\xi(a, a', S'') > \epsilon_c$ instead, which corresponds to Eq. (11b).

(R2): As required in P1, $\Pr[M^{\text{svt}}_{\widetilde{\theta}}(a') \in S'']$ cannot reach the threshold c. So DP-Sniper's empirical $\hat{S}$ includes the entire S'' indiscriminately, and includes b^0 with probability q, i.e.

$$\Pr[b \in \hat{S}] = \mathbb{1}[b \in S''] + q \cdot \mathbb{1}[b = b^0],\ q = \frac{c - \Pr[M^{\text{svt}}_{\widetilde{\theta}}(a') \in S'']}{\Pr[M^{\text{svt}}_{\widetilde{\theta}}(a') = b^0]}.$$

Hence, $\xi^*(\widetilde{\theta}) = \ln(\Pr[M^{\text{svt}}_{\widetilde{\theta}}(a) \in S''] + q \cdot \Pr[M^{\text{svt}}_{\widetilde{\theta}}(a) = b^0]) - \ln(c)$, and the inequality becomes Eq.(11c).

To sum up, any $M^{\text{svt}}_{\widetilde{\theta}}$ satisfying Eq.(11) qualifies as a false positive. It is also a sufficient but not necessary condition for an FP, due to the relaxation applied in our derivation.

The proof of Theorem 3 for benchmark SVT mechanism is as follows. Without loss of generality, we first let $q(a) = 0$ and $q(a') = 1$. It follows that

$$\Pr[M(a) = \perp] = 1 + \frac{1}{6}e^{-\theta/2} - \frac{2}{3}e^{-\theta/4}, \tag{18}$$

which is above $\frac{1}{2}$ for any $\theta \geq 0$. Further, we have $\Pr[M(a') = \perp] = \frac{1}{2}$. Therefore, for any small probability threshold c below $\frac{1}{2}$, we have $\Pr[M(a') = \perp] > c$ and $\Pr[M(a') = \top] > c$, thereby ruling out the false positives following Prerequisite 1. The corresponding power for this adjacent pair is then

$$\ln\left(\frac{\Pr[M(a) = \perp]}{\Pr[M(a') = \perp]}\right) = \ln\left(2 + \frac{1}{3}e^{-\theta/2} - \frac{4}{3}e^{-\theta/4}\right). \tag{19}$$

We then switch a and a', i.e. $q(a) = 1$ and $q(a') = 0$ instead. In this case, with $\Pr[M(a') = \top] < \frac{1}{2}$ being the only possible small output probability, Prerequisite 1 becomes $\Pr[M(a') = \top] < c$, and the corresponding power is

$$\ln\left(\frac{\Pr[M(a) = \top]}{\Pr[M(a') = \top]}\right) = \ln\left(\frac{c}{2}(1 + \frac{c + \frac{1}{6}e^{-\theta/2} - \frac{2}{3}e^{-\theta/4}}{1 + \frac{1}{6}e^{-\theta/2} - \frac{2}{3}e^{-\theta/4}})\right). \tag{20}$$

Summing up the two cases of a and a', the theoretical maximal privacy ϵ^* is Eq. (19), while DP-Sniper's maximal power ξ^* is Eq. (20), thereby finishing the proof.

Proof of Lemma 1 for adapted SVT mechanism is as follows. First, we have

$$\Pr[M(A) = b^j] = \begin{cases} \int_{-\infty}^{\infty} \Pr[\rho = z] g_{j+1}(z) dz,\ j = 0, \\ \int_{-\infty}^{\infty} \Pr[\rho = z] \prod_{i \in [1,j]} f^i_A[z]\ g^{j+1}_A(z) dz,\ j \in [1, N), \\ \int_{-\infty}^{\infty} \Pr[\rho = z] \prod_{i \in [1,j]} f^i_A[z]\ dz,\ j = N, \end{cases} \tag{21}$$

where $f^i_A(z) = \Pr[q_i(A) + v_i < T_i + z]$,
and $g^i_A(z) = \Pr[q_i(A) + v_i \geq T_i + z]$.

We notice that the probabilities of b^0 and $S^0 = \{b^j | 1 \leq j \leq N\}$ are the simplest among all possible outputs, where

$$\begin{aligned} \Pr[M(a) = b^0] &= \frac{1}{\theta_1}\int_{-2\theta_1}^{-\theta_1} (1 - \frac{1}{2}e^{\theta_2(1+z)}) dz, \\ \Pr[M(a') = b^0] &= \frac{1}{\theta_1}\int_{-2\theta_1}^{-\theta_1} (1 - \frac{1}{2}e^{\theta_2 z}) dz. \\ \Pr[M(a) \in S^0] &= 1 - \Pr[M(a) = b^0] = \frac{1}{\theta_1}\int_{-2\theta_1}^{-\theta_1} \frac{1}{2}e^{\theta_2(1+z)} dz, \\ \Pr[M(a') \in S^0] &= 1 - \Pr[M(a') = b^0] = \frac{1}{\theta_1}\int_{-2\theta_1}^{-\theta_1} \frac{1}{2}e^{\theta_2 z} dz. \end{aligned} \tag{22}$$

To prove the validity of Lemma 1, we just need to demonstrate:

$$\forall j \in [1, N],\ r(b^j) > r(b^0), \tag{23}$$

Proof. We will use the following lemma:

Lemma 2. *For countable sequences of fractions $\{\frac{a_n}{b_n}\}$ and $\{\frac{c_n}{d_n}\}$, if for any i holds that $\frac{a_i}{b_i} > \frac{c_i}{d_i}$, then the following inequality holds:*

$$\frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} b_i} > \frac{\sum_{i=1}^{n} c_i}{\sum_{i=1}^{n} d_i}. \tag{24}$$

Now we define the following two Riemann intergrable functions:

$$\begin{aligned} f^j(z) &= (\frac{1}{2}e^{\theta_2(1+z)})^j * (1 - \frac{1}{2}e^{\theta_2(1+z)}), \\ g^j(z) &= (\frac{1}{2}e^{\theta_2 z})^j * (1 - \frac{1}{2}e^{\theta_2 z}). \end{aligned}$$

It is easy to observe that $\frac{f}{g}$ is also Riemann integrable.Note that for any $j \neq 0, \forall z$,

$$
\begin{aligned}
\frac{f^j(z)}{g^j(z)} &= \frac{(\frac{1}{2}e^{\theta_2(1+z)})^j * (1-\frac{1}{2}e^{\theta_2(1+z)})}{(\frac{1}{2}e^{\theta_2 z})^j * (1-\frac{1}{2}e^{\theta_2 z})} \\
&= (e^{\theta_2})^j \frac{(\frac{1}{2}e^{\theta_2 z})^j * (1-\frac{1}{2}e^{\theta_2(1+z)})}{(\frac{1}{2}e^{\theta_2 z})^j * (1-\frac{1}{2}e^{\theta_2 z})} \\
&= (e^{\theta_2})^j \frac{1-\frac{1}{2}e^{\theta_2(1+z)}}{1-\frac{1}{2}e^{\theta_2 z}} \\
&= (e^{\theta_2})^j \frac{f^0(z)}{g^0(z)} \\
&> \frac{f^0(z)}{g^0(z)}.
\end{aligned}
\tag{25}
$$

Divide $[-2\theta_1, -\theta_1]$ into subintervals: $-2\theta_1 = z_1 < z_2 < \cdots < z_n = -\theta_1$, and $\Delta z = |z_i - z_{i-1}| = \frac{\theta_1}{n-1}$. Using the definition of Riemann integration, we obtain the following expression:

$$
\begin{aligned}
\int_{-2\theta_1}^{-\theta_1} f^j(z)\mathrm{d}z &= \lim_{n\to\infty} \sum_{i=1}^{n-1} f^j(z_i)\Delta z, \\
\int_{-2\theta_1}^{-\theta_1} g^j(z)\mathrm{d}z &= \lim_{n\to\infty} \sum_{i=1}^{n-1} g^j(z_i)\Delta z.
\end{aligned}
\tag{26}
$$

Therefore, based on Eq. (25) and Eq. (26), we have obtained a part of conclusion that needs to be proved:

$$
\begin{aligned}
\frac{\Pr[M(a)=b^j]}{\Pr[M(a')=b^j]} &= \frac{\int_{-2\theta_1}^{-\theta_1} f^j(z)\mathrm{d}z}{\int_{-2\theta_1}^{-\theta_1} g^j(z)\mathrm{d}z} \\
&= \lim_{n\to\infty} \frac{\sum_{i=1}^{n-1} f^j(z_i)\Delta z}{\sum_{i=1}^{n-1} g^j(z_i)\Delta z} \\
&> \lim_{n\to\infty} \frac{\sum_{i=1}^{n-1} f^0(z_i)\Delta z}{\sum_{i=1}^{n-1} g^0(z_i)\Delta z} \\
&= \frac{\Pr[M(a)=b^0]}{\Pr[M(a')=b^0]}, \forall j \in [1, N).
\end{aligned}
$$

Finally , for $j = N$, through a simple integration, we easily conclude that the result holds when $j = N$ as well. □

Therefore, b^0 has the lowest ratio among all outputs:

$$
\begin{aligned}
\Pr[M^{\text{svt}}_{\widetilde{\theta}}(a') \in S^0] &= 1 - \Pr[M^{\text{svt}}_{\widetilde{\theta}}(a') \in b^0] \\
&= (e^{-\widetilde{\theta}_1\widetilde{\theta}_2} - e^{-2\widetilde{\theta}_1\widetilde{\theta}_2})/(\widetilde{\theta}_1\widetilde{\theta}_2).
\end{aligned}
\tag{27}
$$

Proof completes.

E EXAMPLE 3: ONE-TIME RAPPOR AGAINST DP-SNIPER AUDITING

Benchmark RAPPOR Mechanism $M^{\text{rappor}}_{\theta}$. The benchmark claims to be ϵ_c-DP and operates as in Alg. 4.

Algorithm 4: One-time RAPPOR mechanism

Input: Dataset A, parameter θ, Bloom filter B of size k, h hash functions
Output: b

Hash A onto the Bloom filter B using h hash functions;
for *each bit $B_i(A)$ in $B(A)$, $0 \leq i < k$* **do**
  $b_i = 1$ with probability $\frac{1}{2}\theta$; 0 with probability $\frac{1}{2}\theta$; $B_i(A)$ with probability $1-\theta$.
end
Output b.

Theorem 10. *The benchmark one-time RAPPOR mechanism M^{rappor}_{θ} is an FP against DP-Sniper auditing with probability threshold c if*

$$
\begin{cases}
(P1)\ \ (\frac{1}{2}\theta)^{2h} < c, & (28a)\\
(R1)\ \ 2h\cdot(\ln(1-\frac{1}{2}\theta) - \ln(\frac{1}{2}\theta)) > \epsilon_c, & (28b)\\
(R2)\ \ (1-\frac{1}{2}\theta)^{2h} + qh\theta(1-\frac{1}{2}\theta)^{2h-1} \leq e^{\epsilon_c}c. & (28c)
\end{cases}
$$

Proof. (P1): We first determine the theoretical S^* to solve for prerequisite 1. Intuitively, an output achieves the largest likelihood ratio when all h bits in $B(a')$ differing from $B(a)$ are flipped. Formally, $\Pr[M^{\text{rappor}}_{\theta}(a') \in S^*] = (\frac{1}{2}\theta)^{2h}$, and P1 becomes Eq. (28a).

(R1): The probability of dataset a on the theoretical optimal outcome set S^* is $\Pr[M^{\text{rappor}}_{\theta}(a) \in S^*] = (1-\frac{1}{2}\theta)^{2h}$. Hence the theoreotical DP level is $\epsilon^*(\theta) = \ln((1-\frac{1}{2}\theta)^{2h}) - \ln((\frac{1}{2}\theta)^{2h})$, and R1 corresponds to Eq. (28b).

(R2): We now identify DP-Sniper's empirical $\hat{S}$. We denote the outcome set with the second largest ratio as S', where $\Pr[M^{\text{rappor}}_{\theta}(a) \in S'] = 2h\cdot(1-\frac{1}{2}\theta)^{2h-1}\cdot(\frac{1}{2}\theta)$ and $\Pr[M^{\text{rappor}}_{\theta}(a') \in S'] = 2h\cdot(1-\frac{1}{2}\theta)\cdot(\frac{1}{2}\theta)^{2h-1}$. To simplify computation, we only consider θ for which $\Pr[M^{\text{rappor}}_{\theta}(a') \in S^* \cup S'] \geq c$, in which case

$$
\Pr[b \in \hat{S}] = \mathbb{1}[b \in S^*] + q\cdot\mathbb{1}[b \in S'],\ q = \frac{c - \Pr[M^{\text{rappor}}_{\widetilde{\theta}}(a') \in S^*]}{\Pr[M^{\text{rappor}}_{\widetilde{\theta}}(a') \in S']}.
$$

Then $\xi^*(\widetilde{\theta}) = \ln(\Pr[M^{\text{rappor}}_{\widetilde{\theta}}(a) \in S^*] + q\cdot\Pr[M^{\text{rappor}}_{\widetilde{\theta}}(a) \in S']) - \ln(c)$, and R2 corresponds to Eq. (28c).

Adapted RAPPOR Mechanism (Omitted). The only parameter involved in the one-time RAPPOR mechanism is the flipping probability f. Hence adapting a false positive mechanism reduces to identifying false positive benchmark RAPPOR, and we omit this redundant discussion.

F PROOF OF PREREQUISITE 2

Given an adapted mechanism M_{ϑ} against MPL's auditing with density threshold τ, the iff condition for $\xi^*(\vartheta) < \epsilon^*(\vartheta)$ is

$$
\forall b \in S^*,\ \min\{p(b|a), p(b|a')\} < \tau. \tag{29}
$$

Proof. First, we establish sufficiency when we satisfy condition $\forall b \in S^*,\ \min\{p(b|a), p(b|a')\} < \tau$. For any $b \in S^*$, we have $\epsilon^*(\vartheta) = \ln(p(b|a)) - \ln(p(b|a'))$. According to the definition of S^*, we can obtain the following inequality $\forall b' \notin S^*, \forall b \in S^*$:

$$\ln(p(b'|a)) - \ln(p(b'|a')) < \ln(p(b|a)) - \ln(p(b|a')) = \epsilon^*(\vartheta). \quad (30)$$

For $\hat{b} \in \hat{S}$, if $\hat{b} \notin S^*$, according to Eq. (30), we have $\xi^*(\vartheta) = \ln(p^{\geq\tau}(\hat{b}|a)) - \ln(p^{\geq\tau}(\hat{b}|a')) \leq \ln(p(\hat{b}|a)) - \ln(p(\hat{b}|a')) < \epsilon^*(\vartheta)$. If $\hat{b} \in S^*$, which satisfies condition $\min\{p(\hat{b}|a), p(\hat{b}|a')\} < \tau$, we have $\xi^*(\vartheta) = \ln(p^{\geq\tau}(\hat{b}|a)) - \ln(p^{\geq\tau}(\hat{b}|a')) = \ln(p(\hat{b}|a)) - \ln\tau < \ln(p(\hat{b}|a)) - \ln(p(\hat{b}|a')) = \epsilon^*(\vartheta)$. In summary, we obtain that $\xi^*(\vartheta) < \epsilon^*(\vartheta)$.

Next, we proceed to prove the necessity. According to the definition of $\xi^*(\vartheta)$, we can conclude that for any b in outcome set, $\xi^* \geq \ln(p^{\geq\tau}(b|a)) - \ln(p^{\geq\tau}(b|a'))$. If there exists $b^* \in S^*$, s.t. $\min(p(b^*|a), p(b^*|a')) \geq \tau$, then $\xi^*(\vartheta) \geq \ln(p^{\geq\tau}(b^*|a)) - \ln(p^{\geq\tau}(b^*|a'))$ $= \ln(p(b^*|a)) - \ln(p(b^*|a')) = \epsilon^*(\vartheta)$. This contradicts the condition. □

G EXAMPLE 3: ONE-TIME RAPPOR AGAINST MPL AUDITING

Benchmark RAPPOR Mechanism M_θ^{rappor}. The detailed derivation is similar to that in §E.

Theorem 11. *The benchmark One-time RAPPOR mechanism M_θ^{rappor} is an FP against MPL auditing with probability threshold τ, if*

$$\begin{cases} (P2) \;\; (\theta/2)^{2h}(1-\theta/2)^{k-2h} < \tau, & (31a) \\ (R1) \;\; (\frac{1-\theta/2}{\theta/2})^{2h} > e^{\epsilon_c}, & (31b) \\ (R2) \;\; \max\{\tau, (1-\theta/2)^k\} \leq e^{\epsilon_c}\tau. & (31c) \end{cases}$$

Proof. (P2): We first determine the theoretical S^* to solve for prerequisite 2. Intuitively, an output achieves the largest likelihood ratio when all h bits in $B(a')$ differing from $B(a)$ are flipped.

Lemma 3. *For bits where $B(a)$ and $B(a')$ are the same, the value of the output on these bits will not affect the likelihood ratio. Furthermore, for all $s^* \in S^*$, we have the following conclusion:*

$$\begin{cases} (1-\frac{1}{2}\theta)^{2h}(\frac{1}{2}\theta)^{k-2h} \leq \Pr(M_\theta^{rappor}(a) = s^*) \leq (1-\frac{1}{2}\theta)^k, \\ (\frac{1}{2}\theta)^k \leq \Pr(M_\theta^{rappor}(a') = s^*) \leq (\frac{1}{2}\theta)^{2h}(1-\frac{1}{2}\theta)^{k-2h}. \end{cases} \quad (32)$$

According to the Lemma 3, we can draw the following conclusion: $\Pr[M_\theta^{\text{rappor}}(a') = s^* \in S^*] \leq (\frac{1}{2}\theta)^{2h}(1-\frac{1}{2}\theta)^{k-2h}$, and P2 becomes Eq. (31a).

(R1): To maximize the likelihood ratio, i.e., for all s^* in S^*, their common characteristic is that the values of the $2h$ bits where $B(a)$ and $B(a')$ differ are the same with those of $B(a)$. For the other $k-2h$ bits, according to the lemma, the values at these positions do not affect the likelihood ratio. Hence the theoreotical DP level is $\epsilon^*(\theta) = \ln((1-\frac{1}{2}\theta)^{2h}) - \ln((\frac{1}{2}\theta)^{2h})$, and R1 corresponds to Eq. (31b).

(R2): We denote all possible outcomes of One-time RAPPOR as S, which includes all binary strings of length k. For the probability of obtaining any particular string s in this set, we have the following probability range:

$$(\frac{1}{2}\theta)^k \leq \Pr[M_\theta^{\text{rappor}}(a)] \leq (1-\frac{1}{2}\theta)^k.$$

This holds true for a' as well. Then $\xi^*(\widetilde{\theta}) \leq \ln(\max((1-\frac{1}{2}\theta)^k, \tau)) - \ln(\tau)$ and R2 corresponds to Eq. (31c). □

H DELTA-SIEGE AUDITING

The Laplace mechanism's theoretical $\beta(\alpha)$ curve and its theoretical $\epsilon - e^\delta$ tradeoff $\mathcal{T}(\epsilon)$ are derived as follows. The former (Eq.15) follows directly from [14]. The latter is the minimal δ not violated by the mechanism's (α, β) pair, which is when the line $\beta = 1-\delta-e^\epsilon\alpha$ tangents the curve

$$\beta(\alpha) = \begin{cases} 1 - e^\theta\alpha, \alpha < \frac{e^{-\theta}}{2}; \\ \frac{1}{4}e^{-\theta}\alpha^{-1}, e^{-\theta}/2 \leq \alpha < \frac{1}{2} \end{cases}$$

in the $\beta - \alpha$ plain. A simple calculation gives us the explicit form of $\mathcal{T}(\epsilon)$. Derivation of the Gaussian mechanism follows similarly.

Next we prove prerequisite 3 as follows.

Case 1: Consider when the auditor's $\rho(\epsilon, \delta)$ contour belongs to a different function class from the mechanism's theoretical $\delta = \mathcal{T}(\epsilon)$. Then even if $c < \Pr[M(a') \in S^*]$ i.e. the auditor can reliably estimate the probabilities at the optimal outcome set, the computed ξ is still different from ϵ^* because of the mismatch between $\rho(\epsilon, \delta)$ and $\mathcal{T}(\epsilon)$, i.e. $\theta^* \neq \xi^*$

Case 2: Consider when $c > \Pr[M(a') \in S^*]$, i.e. the auditor cannot successfully reach the optimal outcome set. Then even if the ρ contour is the same as the theoretical $\mathcal{T}(\epsilon)$, the auditor can only reliably estimates the distinguishability at the inferior witnesses, leaving ξ^* still below ϵ^*.

I DPSGD-AUDITING

Prerequisite 4 (P4). *Given a one-step DP-SGD mechanism M_ϑ against DPSGD-Audit with smallest achievable probability c, $\xi^*(\vartheta) < \epsilon^*(\vartheta)$ iff $\Pr[M_\vartheta(a') \in S^*] < c$, where S^* is the theoretical optimal outcome set in Eq.*(3).

The proof follows directly from [14], with an intuition as illustrated in Fig. 17.

J EXPERIMENTS ON ONE-TIME RAPPOR

Benchmark One-Time RAPPOR Mechanism. For the one-time RAPPOR mechanism in Alg. 4, we let $(k, h) \in \{(6, 3), (8, 3), (12, 3)\}$. The result is shown in Fig. 18 and Fig. 19 for DP-Sniper and MPL, respectively.

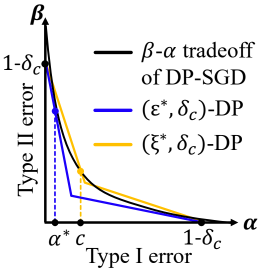
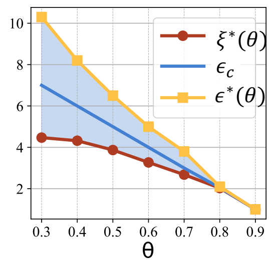


Figure 17: One-step DPSGD against blackbox auditing.

(a) Benchmark RAPPOR against DP-Sniper, $k = 8, h = 3$



(b) Benchmark RAPPOR against DP-Sniper, $k = 12, h = 3$

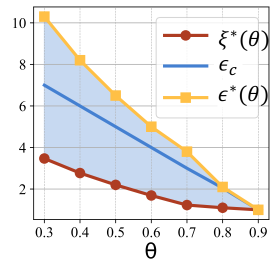
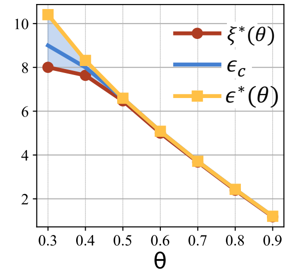


Figure 18: Benchmark RAPPOR mechanism against DP-Sniper's auditing, $c = 0.01$**.**

(a) Benchmark RAPPOR against MPL, $k = 6, h = 3$



(b) Benchmark RAPPOR against MPL, $k = 8, h = 3$

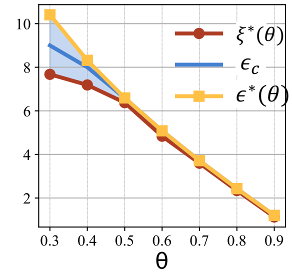


Figure 19: Benchmark One-Time RAPPOR Mechanism against MPL's auditing, $\tau = 10^{-4}$**.**