A. The Edgeworth Approximation

We can apply Edgeworth expansion to approximate $\widetilde{F}_n$ directly, following the techniques introduced in Hall (2013). Let us assume $\boldsymbol{x} \sim Q$. Denote

$$X_Q = \frac{T_n - \mathbb{E}_Q[T_n]}{\sqrt{\mathbb{V}\mathrm{ar}_Q(T_n)}} = \frac{\sum_{i=1}^n (L_i - \mu_i)}{\sqrt{\sum_{i=1}^n \sigma_i^2}}, \tag{A.1}$$

where μ_i and σ_i^2 are the mean and variance of L_i under the distribution Q_i. The characteristic function of X_Q is

$$\chi_n(t) = \exp\left(\sum_{i=1}^{\infty} \tilde{\kappa}_r(X_Q) \frac{(it)^r}{r!}\right),$$

where $\tilde{\kappa}_r(X_Q)$ is the r-th cumulant of X_Q. Details of how to compute the cumulants are summarized in Appendix B. Let $\sigma_n = \sqrt{\sum_{i=1}^n \sigma_i^2}$. Particularly we have

$$\begin{aligned}
\tilde{\kappa}_1(X_Q) &= \mathbb{E}_Q(X_Q) = 0, \\
\tilde{\kappa}_2(X_Q) &= \mathbb{V}\mathrm{ar}_Q(X_Q) = 1, \\
&\vdots \\
\tilde{\kappa}_r(X_Q) &= \tilde{\kappa}_r\left(\sigma_n^{-1} \sum_{i=1}^n (L_i - \mu_i)\right) \\
&= \sigma_n^{-r} \sum_{i=1}^n \tilde{\kappa}_r(L_i), \quad \forall r > 2.
\end{aligned} \tag{A.2}$$

We will denote the sum of n cumulants by $\tilde{\boldsymbol{\kappa}}_r = \sum_{i=1}^n \tilde{\kappa}_r(L_i)$. Under the series expansion of the exponential function, we will have

$$\begin{aligned}
\chi_n(t) &= \exp\left(-\frac{t^2}{2}\right) \exp\left(\sum_{r=3}^{\infty} \frac{\sigma_n^{-r}}{r!} \tilde{\boldsymbol{\kappa}}_r (it)^r\right) \\
&\approx \exp\left(-\frac{t^2}{2}\right) \exp\left(\sum_{r=3,4} \frac{\sigma_n^{-r}}{r!} \tilde{\boldsymbol{\kappa}}_r (it)^r\right) \\
&\approx \exp\left(-\frac{t^2}{2}\right) \Bigg(1 + \sigma_n^{-3} \cdot \overbrace{\frac{1}{6} \tilde{\boldsymbol{\kappa}}_3 (it)^3}^{r_1(it)} \\
&\quad + \sigma_n^{-4} \cdot \overbrace{\frac{1}{24} \tilde{\boldsymbol{\kappa}}_4 (it)^4}^{r_2(it)} + \sigma_n^{-6} \cdot \overbrace{\frac{1}{72} \tilde{\boldsymbol{\kappa}}_3^2 (it)^6}^{r_3(it)}\Bigg).
\end{aligned} \tag{A.3}$$

Since $\chi_n(t) = \int e^{ith} d\widetilde{F}_n(h)$ and $e^{-t^2/2} = \int e^{ith} d\Phi(h)$, we can obtain the corresponding "inverse" expansion:

$$\widetilde{F}_n(h) \approx \Phi(h) + \sigma_n^{-3} \cdot R_1(h) + \sigma_n^{-4} \cdot R_2(h) + \sigma_n^{-6} \cdot R_3(h), \tag{A.4}$$

and $R_j(h)$ is a function whose Fourier-Stieljes transform equals $r_j(it)e^{-t^2/2}$:

$$\int_{-\infty}^{\infty} e^{ith} dR_j(h) = r_j(it) e^{-t^2/2}.$$

Let D denote the differential operator d/dh. We have

$$e^{-t^2/2} = (-it)^{-j} \int_{-\infty}^{\infty} e^{ith} d\left\{D^j \Phi(h)\right\}$$

and hence

$$\int_{-\infty}^{\infty} e^{ith} d\left\{(-D)^j \Phi(h)\right\} = (it)^j e^{-t^2/2}.$$

Let us interpret $r_j(-D)$ as a polynomial in D, we then obtain

$$\int_{-\infty}^{\infty} e^{ith} d\left\{r_j(-D)\Phi(h)\right\} = r_j(it)e^{-t^2/2}.$$

Consequently,

$$R_j(h) = r_j(-D)\phi(h). \tag{A.5}$$

It is well known that for $j \geq 1$,

$$(-D)^j\Phi(h) = -He_{j-1}(h)\phi(h) \tag{A.6}$$

and He_js are the Hermite polynomials:

$$\begin{aligned}
He_0(h) &= 1,\\
He_1(h) &= h,\\
He_2(h) &= h^2 - 1,\\
He_3(h) &= h^3 - 3h,\\
He_4(h) &= h^4 - 6h^2 + 3,\\
He_5(h) &= h^5 - 10h^3 + 15h,\\
He_6(h) &= h^6 - 15h^4 + 45h^2 - 15,\\
He_7(h) &= h^7 - 21h^5 + 105h^3,\\
&\cdots
\end{aligned} \tag{A.7}$$

Combine equations A.4, A.5, A.6 and A.7 we can deduce the final result:

$$\begin{aligned}
\widetilde{F}_n(h) \quad &\approx \quad \Phi(h) + \sigma_n^{-3} \cdot -\frac{1}{6}\tilde{\boldsymbol{\kappa}}_3(h^2-1)\phi(h)\\
&\quad + \sigma_n^{-4} \cdot -\frac{1}{24}\tilde{\boldsymbol{\kappa}}_4(h^3-3h)\phi(h)\\
&\quad + \sigma_n^{-6} \cdot -\frac{1}{72}\tilde{\boldsymbol{\kappa}}_3^2(h^5-10h^3+15h)\phi(h).
\end{aligned} \tag{A.8}$$

In A.3, the truncation happens in both the second and third line. In the second line, we truncated terms where $r \geq 5$. In the following line, we apply the series expansion to the exponential function, and we stopped after taking $t_1 := \sigma_n^{-3} \cdot \frac{1}{6}\tilde{\boldsymbol{\kappa}}_3(it)^3$, $t_2 := \sigma_n^{-4} \cdot \frac{1}{24}\tilde{\boldsymbol{\kappa}}_4(it)^4$ and the square of t_1.

The error stems from truncating $r \geq 5$ terms in the second line will be dominated by $\dfrac{1}{120}\sigma_n^{-5}\tilde{\boldsymbol{\kappa}}_5(it)^5$ in the series expansion. The error stems from truncating the expansion of $r = 3, 4$ terms in the following line will be dominated by the square of t_2: $\sigma_n^{-8} \cdot \frac{1}{576}\tilde{\boldsymbol{\kappa}}_4^2(it)^8$.

Since all L_i's are identically distributed, the cumulants of $L_1, \ldots L_n$ take the same value for any fixed order. Therefore, $\sigma_1 = \cdots = \sigma_n = \sigma$ and $\tilde{\kappa}_r = \tilde{\kappa}_r(L_1) = \cdots = \tilde{\kappa}_r(L_n)$. As a consequence, we have $\sigma_n = \sqrt{n}\sigma$ and $\tilde{\boldsymbol{\kappa}}_r = n\tilde{\kappa}_r$. This leads to

$$\begin{aligned}
&\sigma_n^{-3} \cdot \tilde{\boldsymbol{\kappa}}_3(it)^3 \sim n^{-1/2}(it)^3,\\
&\sigma_n^{-4} \cdot \tilde{\boldsymbol{\kappa}}_4(it)^4 \sim n^{-1}(it)^4,\\
&\sigma_n^{-6} \cdot \tilde{\boldsymbol{\kappa}}_3^2(it)^6 \sim n^{-1}(it)^6,\\
&\sigma_n^{-8} \cdot \tilde{\boldsymbol{\kappa}}_4^2(it)^8 \sim n^{-2}(it)^8,\\
&\sigma_n^{-5} \cdot \tilde{\boldsymbol{\kappa}}_5(it)^5 \sim n^{-3/2}(it)^5.
\end{aligned} \tag{A.9}$$

Hence the error for approximating $\chi_n(t)$ is upper bounded by $O\left(n^{-2}(it)^8 + n^{-3/2}(it)^5\right)$. Next, we connect the characteristic function to CDF $\widetilde{F}_n(h)$. From equations A.5 and A.6, we know the error term will be transformed into $O\left(n^{-2}He_7(h) + n^{-3/2}He_4(h)\right)$ as approximating $\widetilde{F}_n(h)$, which is $O\left(n^{-2}h^7 + n^{-3/2}h^3\right)$.

B. Computing Cumulants From Moments

The cumulants of a random variable X are defined using the cumulant-generating function $K(t)$. It is the natural logarithm of the moment-generating function:

$$K(t) = \log \mathbb{E}\left(e^{tX}\right),$$

and the cumulants are the coefficients in the Taylor expansion of $K(t)$ about the origin:

$$K(t) = \log \mathbb{E}\left(e^{tX}\right) = \sum_{r=0}^{\infty} \kappa_r t^r / r!.$$

For any integer $r \geq 0$, the r-th order non-central moment of X is $\mu_r = \mathbb{E}(X^r)$. Recall the Taylor expansion of the moment-generating function $M(t)$ about the origin

$$M(t) = \mathbb{E}\left(e^{tX}\right) = \sum_{r=0}^{\infty} \mu_r t^r / r! = \exp\left(K(t)\right).$$

The cumulants can be recovered in terms of the moments and vice versa. In general,

$$\kappa_r = \sum_{k=1}^{r} (-1)^{k-1} (k-1)! B_{r,k}(\mu_1, \ldots, \mu_{r-k+1})$$

where $B_{n,k}$ are Bell polynomials. The relationship between the first few cumulants and moments is as the following:

$$\begin{aligned}
\kappa_0 &= 0, \\
\kappa_1 &= \mu_1, \\
\kappa_2 &= \mu_2 - \mu_1^2, \\
\kappa_3 &= \mu_3 - 3\mu_2\mu_1 + 2\mu_1^3, \\
\kappa_4 &= \mu_4 - 4\mu_3\mu_1 - 3\mu_2^2 + 12\mu_2\mu_1^2 - 6\mu_1^4.
\end{aligned}$$

C. $\mathcal{N}(0, 1)$ vs $p\mathcal{N}(\mu, 1) + (1 - p)\mathcal{N}(0, 1))$

Let P be the standard normal distribution $\mathcal{N}(0, 1)$ and Q be a mixture model $p\mathcal{N}(\mu, 1) + (1 - p)\mathcal{N}(0, 1)$ with $\mu \geqslant 0$. We now show that

Lemma C.1.

$$T(P, Q) = pG_\mu + (1 - p)\mathrm{Id}.$$

Proof. The likelihood ratio between Q and P is

$$p\mathrm{e}^{-\frac{1}{2}(x-\mu)^2 + \frac{1}{2}x^2} + 1 - p = p\mathrm{e}^{\mu x - \frac{1}{2}\mu^2} + 1 - p.$$

Since $\mu \geqslant 0$, likelihood ratio tests are thresholding, i.e., $\{x : x > h\}$. The type I and type II errors are

$$\begin{aligned}
\alpha &= P\{x : x > h\} = 1 - \Phi(h), \\
\beta &= Q\{x : x \leqslant h\} \\
&= p\,\mathbb{E}_{x \sim \mathcal{N}(\mu,1)}[1_{\{x : x \leqslant h\}}] + (1 - p)\,\mathbb{E}_{x \sim \mathcal{N}(0,1)}[1_{\{x : x \leqslant h\}}] \\
&= p\Phi(h - \mu) + (1 - p)\Phi(h).
\end{aligned}$$

Inverting the first formula, we have $h = \Phi^{-1}(1 - \alpha)$. So

$$\beta = p\Phi(h - \mu) + (1 - p)\Phi(h) = p\Phi(\Phi^{-1}(1 - \alpha) - \mu) + (1 - p)(1 - \alpha)$$

Making use of the known expression $G_\mu(\alpha) = \Phi\big(\Phi^{-1}(1 - \alpha) - \mu\big)$ and $\mathrm{Id}(\alpha) = 1 - \alpha$, we have

$$T(P, Q)(\alpha) = \beta = pG_\mu(\alpha) + (1 - p)\mathrm{Id}(\alpha).$$

$\square$

D. Details of the Numerical Method

D.1. Proof of Lemma 5.1

Proof. By definition of convex conjugacy, $\delta \geqslant \delta_1(\varepsilon)$ if and only if $f(x) \geqslant 1-\delta-\mathrm{e}^\varepsilon x$ for all $x \in [0,1]$. Since $f = T(P,Q)$ characterizes optimal testing rules, $f(x) \geqslant 1-\delta-\mathrm{e}^\varepsilon x$ for any $x \in [0,1]$ if and only if for any event E, $Q[E] \leqslant \mathrm{e}^\varepsilon P[E]+\delta$. That is,

$$\begin{aligned}\delta_1(\varepsilon) &= \min\{\delta : Q[E] \leqslant \mathrm{e}^\varepsilon P[E] + \delta, \forall E\}\\ &= \max_E Q[E] - \mathrm{e}^\varepsilon P[E]\\ &= \max_E \int_E \big[q(x) - \mathrm{e}^\varepsilon p(x)\big]\mathrm{d}\mu(x).\end{aligned}$$

Obviously, the maximum is attained at the event that the integrand being non-negative. That is, $E = \{x : q(x) - \mathrm{e}^\varepsilon p(x) \geqslant 0\}$. Therefore,

$$\delta_1(\varepsilon) = \int \big(q - \mathrm{e}^\varepsilon p\big)_+ \mathrm{d}\mu.$$

□

D.2. Proof of Lemma 5.2

Proof. By definition of $\otimes$ and Lemma 5.1, we have

$$\begin{aligned}\delta_\otimes(\varepsilon) &= 1 + (f_1 \otimes f_2)^*(-\mathrm{e}^\varepsilon)\\ &= 1 + \big(T(P_1 \times P_2, Q_1 \times Q_2)\big)^*(-\mathrm{e}^\varepsilon) && \text{(Def of } \otimes)\\ &= \iint \big(q_1(x)q_2(y) - \mathrm{e}^\varepsilon p_1(x)p_2(y)\big)_+ \mathrm{d}x\mathrm{d}y && \text{(Lemma 5.1)}\\ &= \iint q_2(y)\cdot\Big(q_1(x) - \mathrm{e}^\varepsilon p_1(x)\cdot \tfrac{p_2(y)}{q_2(y)}\Big)_+ \mathrm{d}x\mathrm{d}y && (q_2(y) \geqslant 0)\\ &= \iint q_2(y)\cdot\Big(q_1(x) - \mathrm{e}^{\varepsilon - L_2(y)} p_1(x)\Big)_+ \mathrm{d}x\mathrm{d}y && \text{(Def of } L_2)\\ &= \int q_2(y)\cdot\Big[\int \big(q_1(x) - \mathrm{e}^{\varepsilon - L_2(y)} p_1(x)\big)_+ \mathrm{d}x\Big]\mathrm{d}y && \text{(Fubini)}\\ &= \int q_2(y)\cdot\delta_1\big(\varepsilon - L_2(y)\big)\mathrm{d}y. && \text{(Lemma 5.1 on } \delta_1)\end{aligned}$$

□

E. Privacy Guarantees for Noisy SGD with Sampling Rate $p = \frac{0.5}{\sqrt{n}}$

In Section 5.3 we present the result when the sampling rate $p = 0.5/n^{\frac{1}{4}}$. Since the convergence of `CLT` requires the assumption $p\sqrt{n} \to \nu > 0$ (Bu et al., 2019), that is a regime where the performance of `CLT` does not have theoretical guarantees. Here we present the results when $p = 0.5/n^{\frac{1}{2}}$, where the convergence of `CLT` is guaranteed. However, we still observe that `Edgeworth` outperforms `CLT` . See Figure E.1 and E.2 for the comparison.

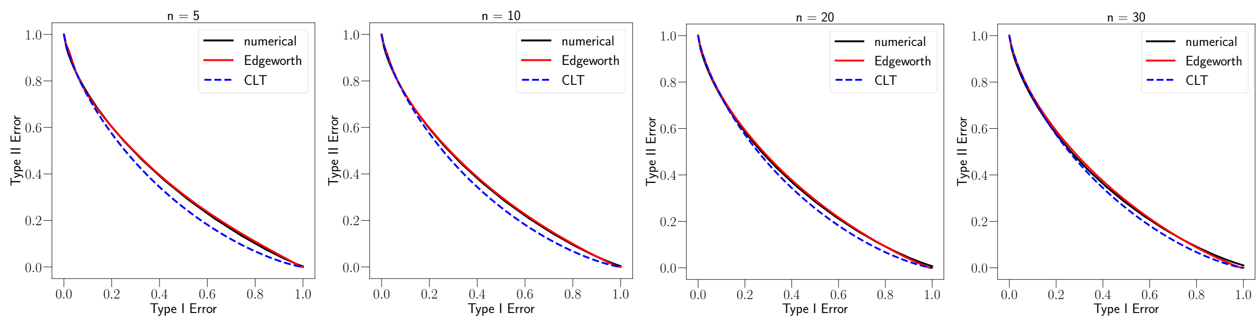

Figure E.1. The estimation of $0.5/n^{\frac{1}{2}}(G_1 + \mathrm{Id})^{\otimes n}$.

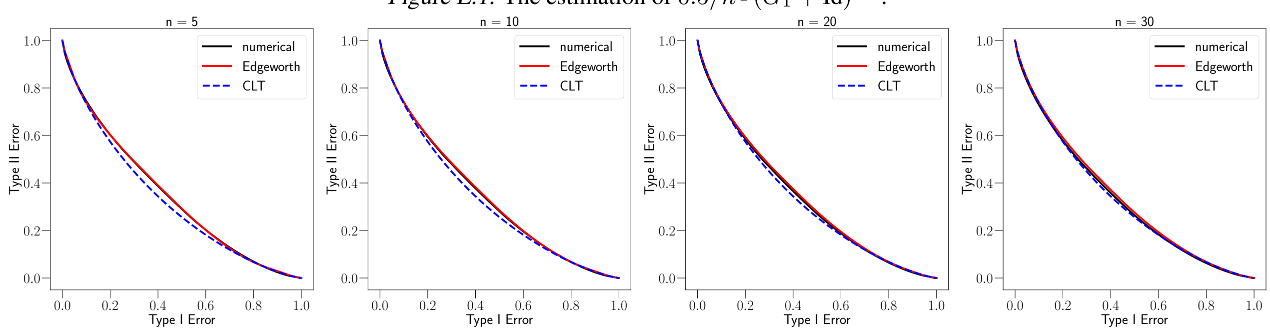

Figure E.2. The estimation of the privacy bound for n-step noisy SGD. The sampling rate is $p = 0.5/n^{\frac{1}{2}}$ and the noise scale is $\sigma = 1$.