Gradient Ascent-based Mean Shift for Differentially Private Nonparametric Mode Estimation: Supplementary Materials

1 Additional Simulations and Results

The codes to reproduce all the application results in the main paper and in this supplementary can be found in here.

1.1 Bivariate 4-Modal Gaussian Mixture

Table 1: MSE and runtime comparison for the bivariate 4-modal Gaussian mixture. Reported are mean ± standard error over repeated runs.

n	ϵ	MSE		Runtime (s)	
		DP-GRAMS	MS	DP-GRAMS	MS
700	0.1	6.793 ± 1.470	0.00897 ± 0.00043	0.00455 ± 0.00015	0.00968 ± 0.00004
	0.2	4.188 ± 0.654	0.00897 ± 0.00043	0.00473 ± 0.00019	0.00968 ± 0.00004
	0.5	0.881 ± 0.120	0.00897 ± 0.00043	0.00449 ± 0.00014	0.00968 ± 0.00004
	1.0	0.231 ± 0.022	0.00897 ± 0.00043	0.00420 ± 0.00011	0.00968 ± 0.00004
	5.0	0.034 ± 0.003	0.00897 ± 0.00043	0.00435 ± 0.00013	0.00968 ± 0.00004
1000	0.1	3.792 ± 0.647	0.00592 ± 0.00011	0.00756 ± 0.00022	0.01808 ± 0.00024
	0.2	1.707 ± 0.285	0.00592 ± 0.00011	0.00741 ± 0.00015	0.01808 ± 0.00024
	0.5	0.374 ± 0.033	0.00592 ± 0.00011	0.00732 ± 0.00021	0.01808 ± 0.00024
	1.0	0.093 ± 0.011	0.00592 ± 0.00011	0.00737 ± 0.00018	0.01808 ± 0.00024
	5.0	0.016 ± 0.002	0.00592 ± 0.00011	0.00721 ± 0.00020	0.01808 ± 0.00024
2000	0.1	3.253 ± 0.435	0.00611 ± 0.00009	0.01455 ± 0.00038	0.07111 ± 0.00026
	0.2	1.065 ± 0.121	0.00611 ± 0.00009	0.01369 ± 0.00029	0.07111 ± 0.00026
	0.5	0.131 ± 0.018	0.00611 ± 0.00009	0.01431 ± 0.00032	0.07111 ± 0.00026
	1.0	0.044 ± 0.005	0.00611 ± 0.00009	0.01386 ± 0.00030	0.07111 ± 0.00026
	5.0	0.010 ± 0.001	0.00611 ± 0.00009	0.01441 ± 0.00031	0.07111 ± 0.00026
5000	0.1	0.649 ± 0.079	0.00195 ± 0.00000	0.07293 ± 0.00084	0.45564 ± 0.00105
	0.2	0.242 ± 0.025	0.00195 ± 0.00000	0.07193 ± 0.00113	0.45564 ± 0.00105
	0.5	0.030 ± 0.003	0.00195 ± 0.00000	0.07439 ± 0.00104	0.45564 ± 0.00105
	1.0	0.014 ± 0.002	0.00195 ± 0.00000	0.07208 ± 0.00077	0.45564 ± 0.00105
	5.0	0.003 ± 0.000	0.00195 ± 0.00000	0.07360 ± 0.00086	0.45564 ± 0.00105

Figure 1 presents a grid of 20 independent algorithm runs on a single realization of the 4-modal Gaussian mixture; each subplot shows the data with the true modes, non-private MS estimates, and DP-GRAMS estimates. Table 1 reports mean ± standard error for MSE and runtimes across sample sizes and privacy budgets.

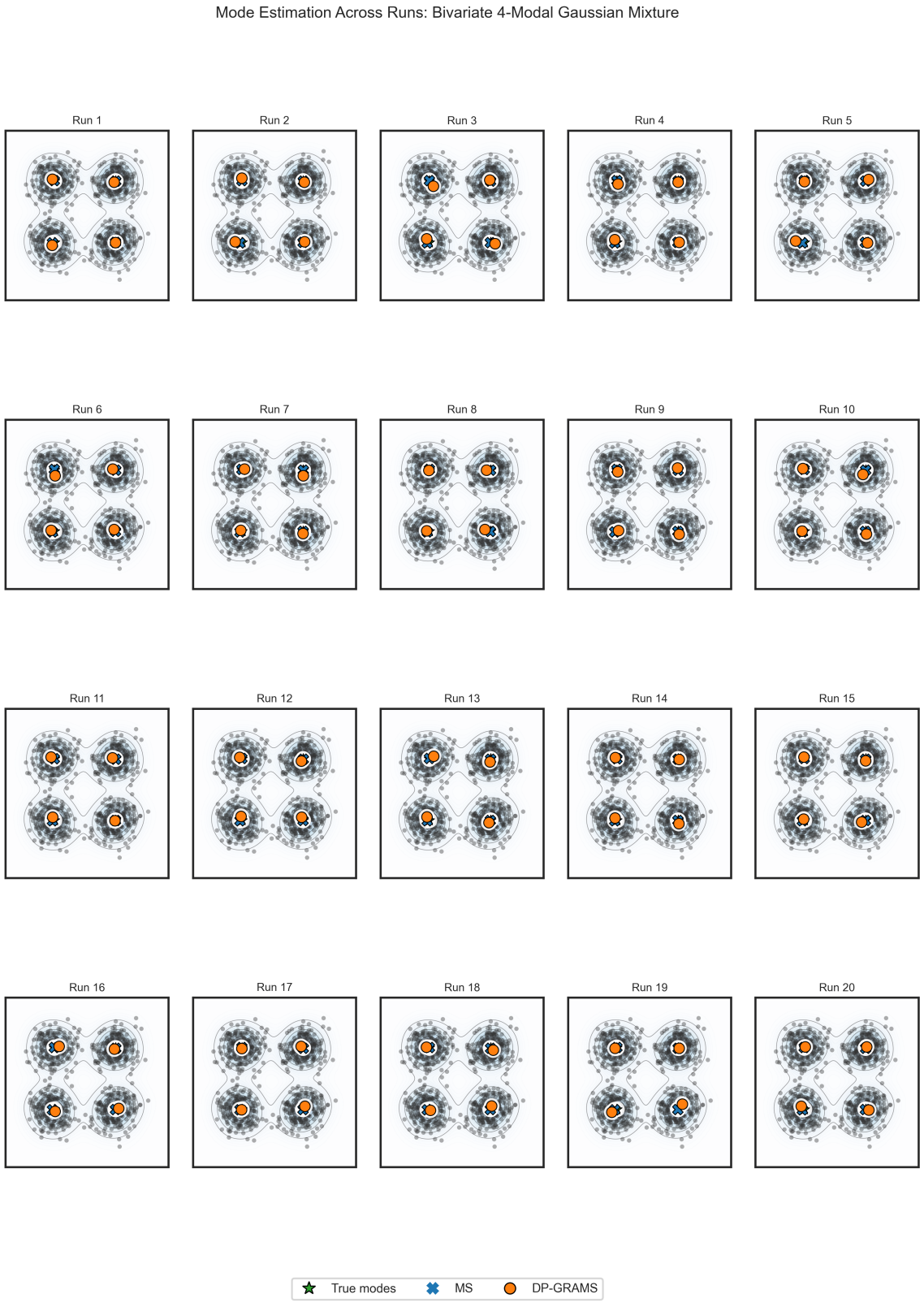


Figure 1: Grid of 20 simulation runs for the bivariate 4-modal Gaussian mixture. Each subplot presents the KDE contour of a single run, with true modes, MS estimates, and DP-GRAMS estimates indicated by green stars, blue crosses, and orange circles, respectively. This grid allows visual assessment of variability in mode estimation across multiple trials and highlights the consistency of both the non-private and differentially private methods under repeated sampling.

1.2 Bivariate 5-Modal t-Mixture

Table 2: MSE and runtime comparison for the bivariate 5-modal t-mixture. Reported are mean $\pm$ standard error over repeated runs.

n	ϵ	MSE		Runtime (s)	
		DP-GRAMS	MS	DP-GRAMS	MS
700	0.1	3.748 ± 0.852	0.00451 ± 0.00004	0.01081 ± 0.00010	0.00968 ± 0.00002
	0.2	2.041 ± 0.216	0.00451 ± 0.00004	0.01055 ± 0.00008	0.00968 ± 0.00002
	0.5	0.452 ± 0.050	0.00451 ± 0.00004	0.01050 ± 0.00009	0.00968 ± 0.00002
	1.0	0.129 ± 0.014	0.00451 ± 0.00004	0.01086 ± 0.00014	0.00968 ± 0.00002
	5.0	0.021 ± 0.003	0.00451 ± 0.00004	0.01103 ± 0.00011	0.00968 ± 0.00002
1000	0.1	2.265 ± 0.283	0.01183 ± 0.00017	0.01556 ± 0.00016	0.01745 ± 0.00001
	0.2	1.104 ± 0.110	0.01183 ± 0.00017	0.01556 ± 0.00017	0.01745 ± 0.00001
	0.5	0.224 ± 0.024	0.01183 ± 0.00017	0.01523 ± 0.00013	0.01745 ± 0.00001
	1.0	0.076 ± 0.007	0.01183 ± 0.00017	0.01530 ± 0.00013	0.01745 ± 0.00001
	5.0	0.016 ± 0.002	0.01183 ± 0.00017	0.01520 ± 0.00019	0.01745 ± 0.00001
2000	0.1	1.367 ± 0.191	0.00330 ± 0.00002	0.04065 ± 0.00047	0.06968 ± 0.00008
	0.2	0.531 ± 0.051	0.00330 ± 0.00002	0.03952 ± 0.00033	0.06968 ± 0.00008
	0.5	0.103 ± 0.010	0.00330 ± 0.00002	0.03884 ± 0.00039	0.06968 ± 0.00008
	1.0	0.028 ± 0.002	0.00330 ± 0.00002	0.04002 ± 0.00045	0.06968 ± 0.00008
	5.0	0.006 ± 0.000	0.00330 ± 0.00002	0.04004 ± 0.00045	0.06968 ± 0.00008
5000	0.1	0.469 ± 0.037	0.00292 ± 0.00002	0.14406 ± 0.00129	0.45242 ± 0.00326
	0.2	0.138 ± 0.015	0.00292 ± 0.00002	0.14154 ± 0.00122	0.45242 ± 0.00326
	0.5	0.029 ± 0.003	0.00292 ± 0.00002	0.14096 ± 0.00082	0.45242 ± 0.00326
	1.0	0.009 ± 0.001	0.00292 ± 0.00002	0.14141 ± 0.00092	0.45242 ± 0.00326
	5.0	0.004 ± 0.000	0.00292 ± 0.00002	0.14304 ± 0.00120	0.45242 ± 0.00326

To probe performance under heavy tails, Figure 2 shows 20 repeated runs for the 5-modal t-mixture, visualizing the data, MS modes and DP-GRAMS-estimated modes. Complementary numerical summaries (mean $\pm$ SE) for MSE and runtime are given in Table 2. The experimental parameters used match those in the main paper.

1.3 Private Modal Regression: 3-Component Mixture

This subsection provides extended numerical and visual evidence for DP-PMS. Figure 3 compares PMS and LOWESS baselines, shows a representative DP-PMS run at $\varepsilon = 1$, and plots the privacy–utility tradeoff across sample sizes. Table 3 lists mean $\pm$ SD for DP-MSE and DP-runtime; these complement the main text by showing how error and runtime scale with n and ε (results averaged over $n_{\text{runs}} = 20$).

1.4 Private Modal Regression: Sinusoidal 2-Mixture Data

We include an additional non-linear test case to demonstrate DP-PMS on smoothly varying modal structure. The sinusoidal 2-mixture dataset is generated as follows. For n samples, let $n_1 = \lfloor n/2 \rfloor$ and $n_2 = n - n_1$, and draw

$$X_1 \sim \text{Uniform}(0,1)^{n_1}, \quad X_2 \sim \text{Uniform}(0,1)^{n_2}.$$

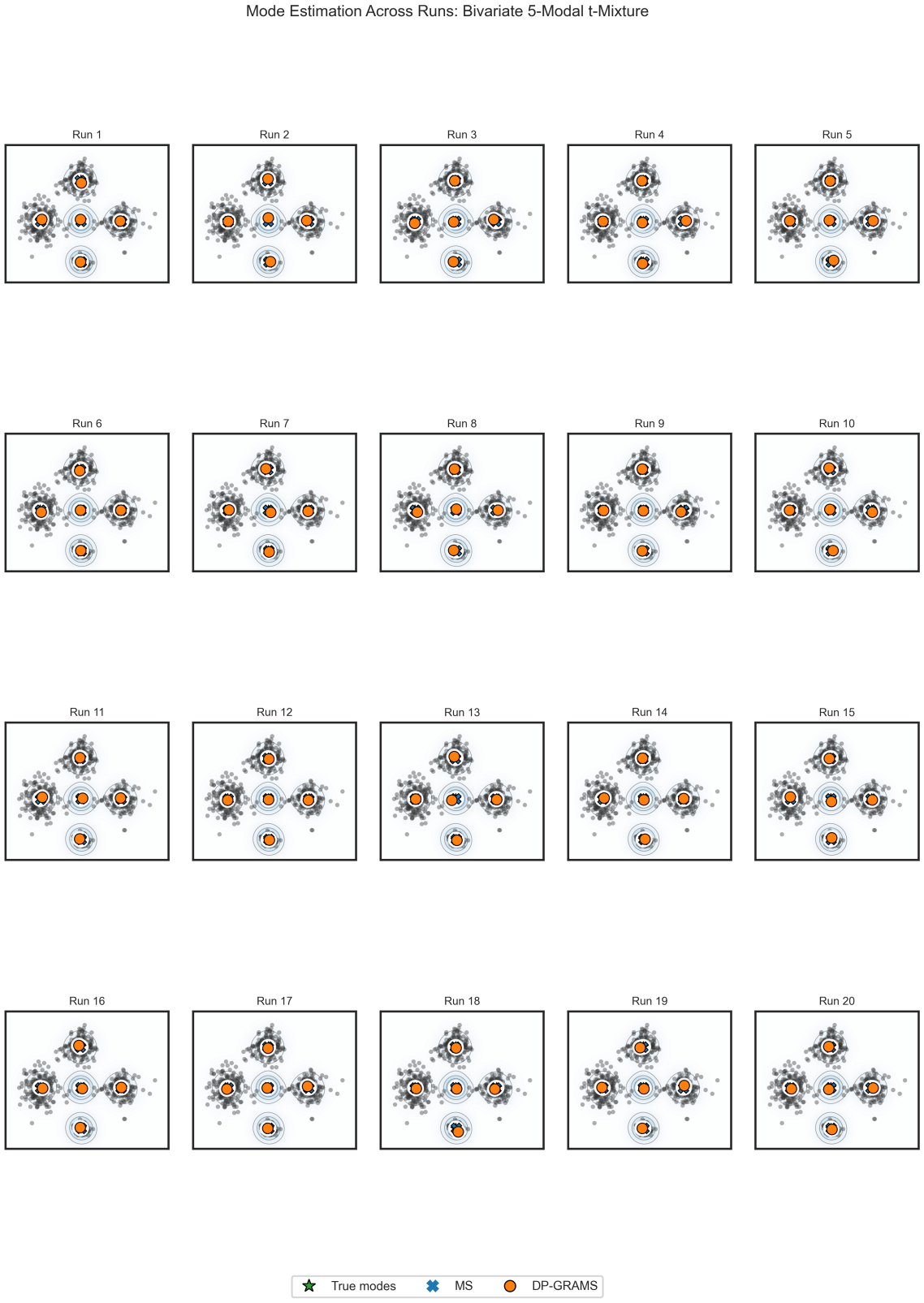


Figure 2: Grid of 20 simulation runs for the bivariate 5-modal t-mixture. Each subplot shows the KDE contours of a single run with the overlaid data points and estimated modes. True modes are indicated by green stars, mean-shift estimates by blue crosses, and DP-GRAMS estimates by orange circles. This layout allows visual assessment of variability in mode estimation across trials and demonstrates the consistency of mean-shift as well as the influence of privacy noise on DP-GRAMS outputs.

(a) PMS vs. LOWESS. (b) DP-PMS at $\varepsilon = 1$. (c) Privacy–utility (MSE vs. ε).

Figure 3: **Modal regression in 3-component mixture data.** (a) Baseline comparison of Partial Mean Shift (PMS) with a nonparametric LOWESS smoother. (b) Differentially private PMS (DP-PMS) output under privacy budget $\varepsilon = 1$. (c) Privacy–utility trade-off for DP-PMS across sample sizes n and privacy budgets ε, showing decreasing error with larger n and ε.

Table 3: Private Modal Regression on 3-Component Mixture Data (mean ± SD for DP-MSE and DP-runtime)

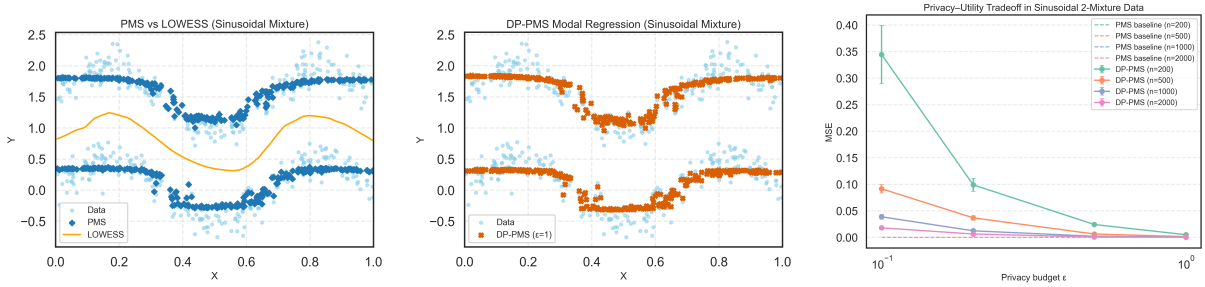
n_{samples}	ε	DP-MSE	PMS-runtime	DP-runtime
300	0.1	0.2461 ± 0.0354	0.010	0.0355 ± 0.0072
300	0.2	0.0526 ± 0.0064	0.010	0.0346 ± 0.0076
300	0.5	0.00505 ± 0.00046	0.010	0.0333 ± 0.0054
300	1.0	0.00227 ± 0.00029	0.010	0.0312 ± 0.0063
600	0.1	0.1021 ± 0.0110	0.020	0.1017 ± 0.0182
600	0.2	0.0185 ± 0.0024	0.020	0.1013 ± 0.0137
600	0.5	0.00216 ± 0.00029	0.020	0.1016 ± 0.0153
600	1.0	0.00069 ± 0.00010	0.020	0.0996 ± 0.0207
1200	0.1	0.0273 ± 0.0030	0.060	0.3104 ± 0.0542
1200	0.2	0.00881 ± 0.00111	0.060	0.3085 ± 0.0579
1200	0.5	0.00127 ± 0.00014	0.060	0.2965 ± 0.0503
1200	1.0	0.00042 ± 0.00009	0.060	0.3124 ± 0.0538
2100	0.1	0.01138 ± 0.00115	0.200	0.9074 ± 0.2030
2100	0.2	0.00284 ± 0.00031	0.200	0.8138 ± 0.1315
2100	0.5	0.00064 ± 0.00007	0.200	0.8211 ± 0.1454
2100	1.0	0.00017 ± 0.00002	0.200	0.8784 ± 0.1618

Then generate the responses with Gaussian noise $\sigma = 0.15$:

$$Y_1 = 1.5 + 0.5\sin(3\pi X_1) + \mathcal{N}(0, \sigma^2), \quad Y_2 = 0.5\sin(3\pi X_2) + \mathcal{N}(0, \sigma^2),$$

and concatenate to form the dataset

$$X = [X_1, X_2], \quad Y = [Y_1, Y_2].$$



(a) PMS vs. LOWESS. (b) DP-PMS at $\varepsilon = 1$. (c) Privacy–utility (MSE vs. ε).

Figure 4: **Modal regression on sinusoidal 2-component mixture data.** (a) Baseline comparison of Partial Mean Shift (PMS) with a smooth nonparametric LOWESS baseline. (b) Differentially private PMS (DP-PMS) output under privacy budget $\varepsilon = 1$. (c) Privacy–utility trade-off for DP-PMS across sample sizes n and privacy budgets ε, showing decreasing error with larger n and ε.

Table 4: Private Modal Regression on Sinusoidal 2-Component Mixture Data (mean ± SD for DP-MSE and DP-runtime)

n_{samples}	ε	DP-MSE	PMS-runtime	DP-runtime
200	0.1	0.3444 ± 0.0547	0.0066	0.0217 ± 0.0041
200	0.2	0.0990 ± 0.0121	0.0066	0.0213 ± 0.0050
200	0.5	0.0243 ± 0.0021	0.0066	0.0203 ± 0.0046
200	1.0	0.0051 ± 0.0005	0.0066	0.0206 ± 0.0046
500	0.1	0.0916 ± 0.0077	0.0230	0.0776 ± 0.0116
500	0.2	0.0368 ± 0.0042	0.0230	0.0781 ± 0.0168
500	0.5	0.0064 ± 0.0007	0.0230	0.0773 ± 0.0110
500	1.0	0.0015 ± 0.0002	0.0230	0.0734 ± 0.0121
1000	0.1	0.0388 ± 0.0045	0.0592	0.2184 ± 0.0314
1000	0.2	0.0124 ± 0.0013	0.0592	0.2072 ± 0.0420
1000	0.5	0.0021 ± 0.0002	0.0592	0.2123 ± 0.0387
1000	1.0	0.0005 ± 0.0000	0.0592	0.2091 ± 0.0355
2000	0.1	0.0180 ± 0.0021	0.1975	0.9180 ± 0.1360
2000	0.2	0.0064 ± 0.0007	0.1975	1.6768 ± 0.2925
2000	0.5	0.0008 ± 0.0001	0.1975	1.6536 ± 0.2972
2000	1.0	0.0002 ± 0.0000	0.1975	1.6716 ± 0.3012

Figure 4 shows PMS vs. DP-PMS for a representative run, and presents the MSE vs. ε curves for multiple n. Numerical summaries appear in Table 4; together these illustrate that DP-PMS closely follows the non-private baseline when privacy noise is small and that larger samples reduce error.

1.5 Private Clustering

This subsection aggregates extended clustering results on Iris, Digits, and MNIST. Per-ε comparisons between DP-GRAMS-C and DP-KMeans (ARI, NMI, MSE, runtime) appear in Tables 5–7; Table 8 gives a compact summary at $(\varepsilon, \delta) = (1, 10^{-5})$. PCA visualizations (Figures 5, 7, 9) and the 1×3 privacy–utility panels (Figures 6 , 8, 10,) make the comparisons intuitive:

- **Iris:** DP-GRAMS-C recovers cluster structure and center estimates comparable to non-private MS and standard k-means (see Table 8). The MSE, ARI, and NMI columns show DP-GRAMS-C offers a favorable utility–privacy tradeoff to DP-kmeans (see Table 5).
- **Digits:** DP-GRAMS-C maintains strong ARI/NMI relative to the non-private MS baseline, whereas DP-KMeans often yields degraded similarity and inflated MSE (see Table 6). PCA plots in Figure 7 visually explain these differences: DP-GRAMS-C preserves multi-modal geometry that k-means is ill-suited to capture.
- **MNIST:** On this large, high-dimensional benchmark (PCA reduced to 50 components), DP-GRAMS-C continues to deliver competitive ARI/NMI with reasonable runtimes, while DP-KMeans shows larger degradation in center MSE for the same privacy budget (see Table 7 and Figure 9).

Collectively these real-data results corroborate the main paper claim: DP-GRAMS-C is a practical private mode-based clustering tool that retains geometric fidelity of clusters (especially for irregular shapes) while providing rigorous (ε, δ) guarantees.

Table 5: Clustering Results on Iris Dataset (mean ± SD for ARI, NMI, MSE, and Runtime)

ε	DP-GRAMS-C				DP-KMeans			
	ARI	NMI	MSE	Runtime (s)	ARI	NMI	MSE	Runtime (s)
0.1	0.6458 ± 0.0747	0.7146 ± 0.0563	3.0811 ± 1.8396	0.0188 ± 0.0046	0.4852 ± 0.1678	0.5697 ± 0.1668	6.7370 ± 1.8159	0.0085 ± 0.0007
0.5	0.7337 ± 0.0435	0.7585 ± 0.0298	0.1926 ± 0.0702	0.0174 ± 0.0001	0.4566 ± 0.2252	0.5202 ± 0.2304	6.5284 ± 2.0236	0.0084 ± 0.0005
1	0.7423 ± 0.0384	0.7529 ± 0.0278	0.1477 ± 0.0414	0.0173 ± 0.0001	0.5020 ± 0.1460	0.5888 ± 0.1621	5.5344 ± 1.7546	0.0080 ± 0.0007
2	0.7219 ± 0.0406	0.7397 ± 0.0283	0.1410 ± 0.0463	0.0175 ± 0.0002	0.5319 ± 0.1118	0.6183 ± 0.1051	4.0118 ± 1.7508	0.0079 ± 0.0007
5	0.7135 ± 0.0285	0.7260 ± 0.0176	0.1275 ± 0.0361	0.0176 ± 0.0002	0.5863 ± 0.0979	0.6775 ± 0.0720	2.2271 ± 1.5790	0.0076 ± 0.0008
10	0.7302 ± 0.0378	0.7411 ± 0.0266	0.1320 ± 0.0361	0.0174 ± 0.0001	0.6526 ± 0.0953	0.7079 ± 0.0619	0.9475 ± 1.1842	0.0075 ± 0.0008

Table 6: Clustering Results on Digits Dataset (mean ± SD for ARI, NMI, MSE, and Runtime)

ε	DP-GRAMS-C				DP-KMeans			
	ARI	NMI	MSE	Runtime (s)	ARI	NMI	MSE	Runtime (s)
0.1	0.7107 ± 0.0000	0.7861 ± 0.0000	5.0648 ± 0.0000	1.0796 ± 0.0158	0.0816 ± 0.0299	0.1934 ± 0.0561	2235.3805 ± 283.9663	0.3898 ± 0.0277
0.5	0.7107 ± 0.0000	0.7861 ± 0.0000	5.0648 ± 0.0000	1.0832 ± 0.0019	0.0361 ± 0.0349	0.1272 ± 0.0731	2029.7903 ± 389.0099	0.3712 ± 0.0304
1	0.7107 ± 0.0000	0.7861 ± 0.0000	5.0648 ± 0.0000	1.0829 ± 0.0040	0.0200 ± 0.0345	0.0685 ± 0.0837	2060.2696 ± 304.4300	0.3278 ± 0.0400
2	0.7107 ± 0.0000	0.7861 ± 0.0000	5.0648 ± 0.0000	1.0806 ± 0.0017	0.0620 ± 0.0587	0.1635 ± 0.0989	1682.0028 ± 264.1345	0.3003 ± 0.0337
5	0.7107 ± 0.0000	0.7861 ± 0.0000	5.0648 ± 0.0000	1.1067 ± 0.0155	0.1867 ± 0.0503	0.3549 ± 0.0595	1216.9286 ± 292.6117	0.2954 ± 0.0207
10	0.7107 ± 0.0000	0.7861 ± 0.0000	5.0648 ± 0.0000	1.1095 ± 0.0052	0.2851 ± 0.0599	0.4562 ± 0.0643	924.1987 ± 281.1311	0.3012 ± 0.0289

Table 7: Clustering Results on MNIST Dataset (mean ± SD for ARI, NMI, MSE, and Runtime)

ε	DP-GRAMS-C				DP-KMeans			
	ARI	NMI	MSE	Runtime (s)	ARI	NMI	MSE	Runtime (s)
0.1	0.4085 ± 0.0321	0.5374 ± 0.0202	45.8043 ± 12.3477	616.49 ± 139.23	0.0244 ± 0.0283	0.0641 ± 0.0586	1070.21 ± 144.57	1.3961 ± 0.2612
0.5	0.4085 ± 0.0321	0.5374 ± 0.0202	45.8043 ± 12.3477	612.06 ± 121.86	0.1795 ± 0.0531	0.2993 ± 0.0593	501.14 ± 143.66	1.9481 ± 0.2595
1	0.4085 ± 0.0321	0.5374 ± 0.0202	45.8043 ± 12.3477	654.77 ± 191.25	0.2374 ± 0.0573	0.3642 ± 0.0473	264.29 ± 121.63	1.3860 ± 0.1436
2	0.4085 ± 0.0321	0.5374 ± 0.0202	45.8043 ± 12.3477	567.48 ± 122.44	0.2506 ± 0.0541	0.3794 ± 0.0434	153.61 ± 94.33	1.3890 ± 0.1378
5	0.4085 ± 0.0321	0.5374 ± 0.0202	45.8043 ± 12.3477	614.93 ± 149.06	0.2569 ± 0.0502	0.3871 ± 0.0376	94.39 ± 40.40	1.3597 ± 0.1045
10	0.4085 ± 0.0321	0.5374 ± 0.0202	45.8043 ± 12.3477	564.39 ± 117.71	0.2584 ± 0.0489	0.3896 ± 0.0360	75.28 ± 19.79	1.3601 ± 0.1474

Table 8: Clustering performance on real-world datasets. Privacy parameters for DP-GRAMS-C and DP-KMeans are $(\varepsilon, \delta) = (1, 10^{-5})$. MSE denotes mean-squared error to the true cluster centers.

Dataset	Samples	Features	True Clusters	Algorithm	ARI	NMI	MSE	Runtime (s)
Iris	150	4	3	MS Clustering	0.784	0.761	0.222	0.009
				DP-GRAMS-C	0.702	0.734	0.992	0.025
				KMeans	0.730	0.758	0.066	0.028
				DP-KMeans	0.438	0.547	5.138	0.012
Digits	1797	64	10	MS Clustering	0.711	0.785	5.113	2.797
				DP-GRAMS-C	0.711	0.786	5.065	1.090
				KMeans	0.669	0.747	3.917	0.084
				DP-KMeans	0.000	0.006	2247.698	0.351
MNIST	70,000	784	10	MS Clustering	0.446	0.569	37.631	467.396
				DP-GRAMS-C	0.423	0.545	40.501	181.159
				KMeans	0.366	0.485	39.475	1.153
				DP-KMeans	0.243	0.358	313.218	0.642

Figure 5: Iris dataset: PCA 2D clustering visualization and comparison across 4 clustering methods: MS (with Agglomerative Merging post-processing step), DP-GRAMS-C (with Agglomerative Merging post-processing step), KMeans and DP K-Means.

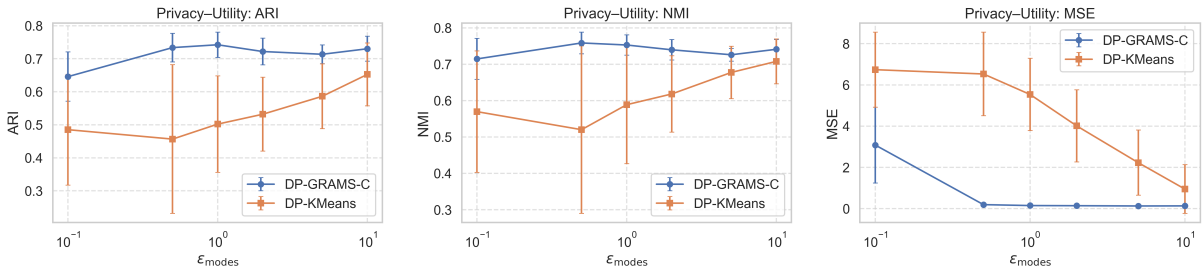


Figure 6: Privacy-utility plots for Iris dataset: ARI, NMI, and MSE.

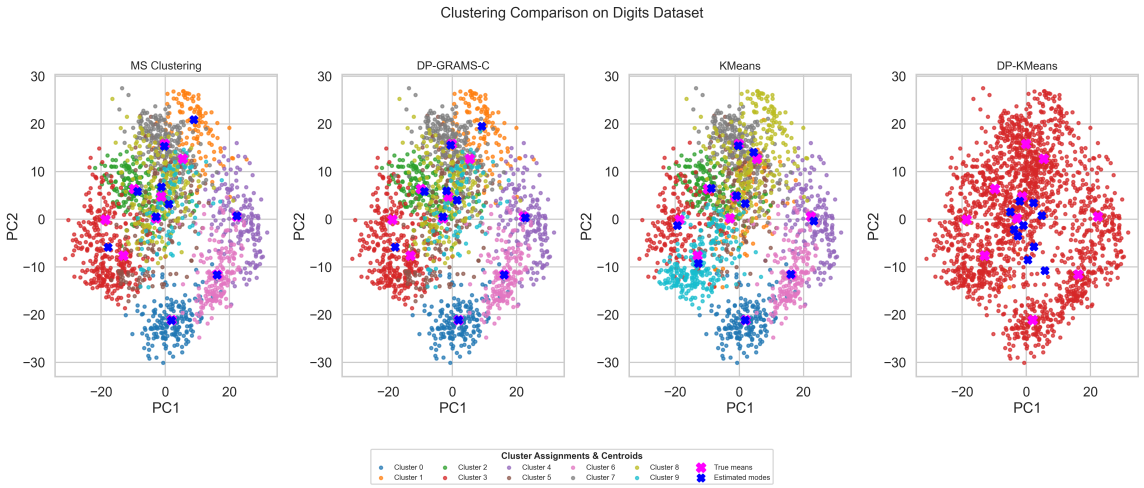


Figure 7: Digits dataset: PCA 2D clustering visualization and comparison across 4 clustering methods: MS (with Agglomerative Merging post-processing step), DP-GRAMS-C (with Agglomerative Merging post-processing step), KMeans and DP K-Means.

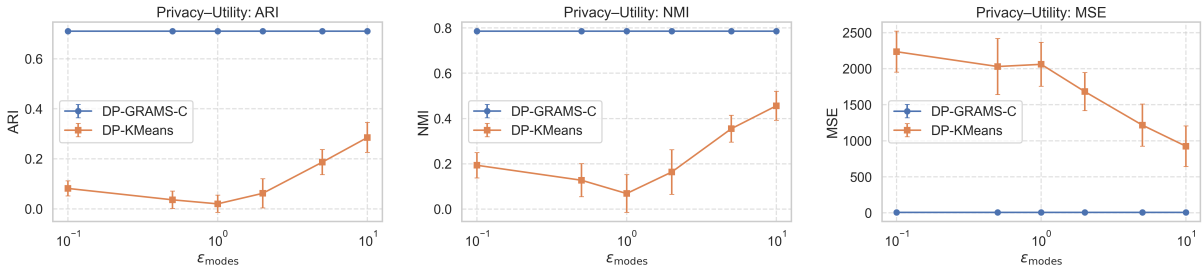


Figure 8: Privacy-utility plots for Digits dataset: ARI, NMI, and MSE.

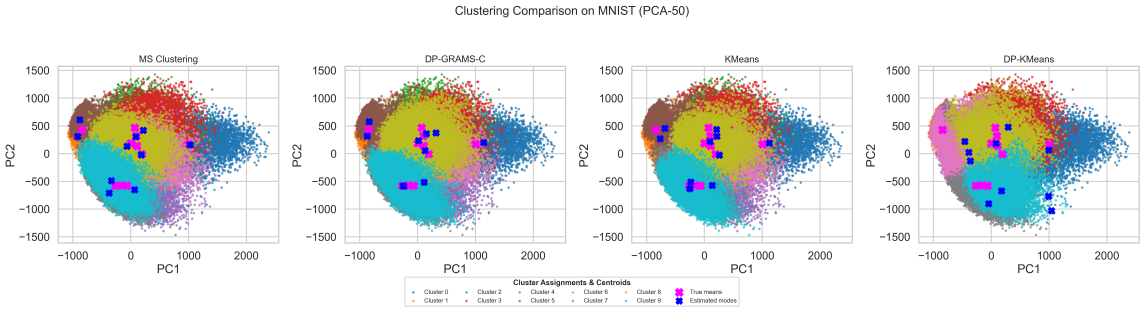


Figure 9: MNIST dataset (PCA to 50 components): clustering visualization and comparison across 4 clustering methods: MS (with Agglomerative Merging post-processing step), DP-GRAMS-C (with Agglomerative Merging post-processing step), KMeans and DP K-Means.

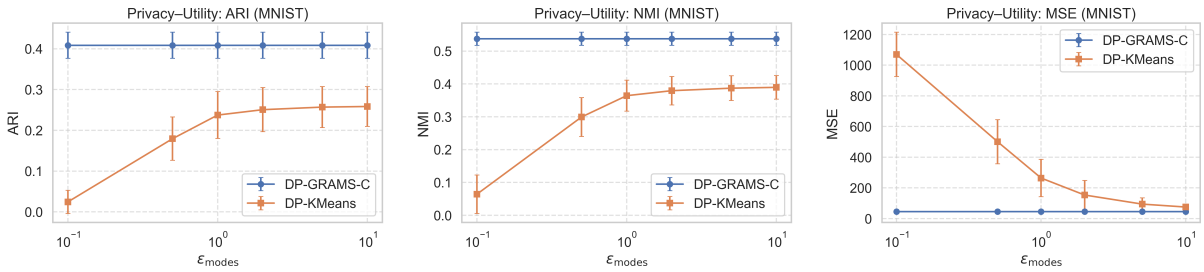


Figure 10: Privacy-utility plots for MNIST dataset: ARI, NMI, and MSE.

2 Algorithms

In this section, we provide detailed pseudocode for the differentially private algorithms used throughout the paper. These algorithms complement the high-level descriptions in the main text and clarify the procedural steps, parameter choices, and post-processing operations.

Algorithm 1: DP-GRAMS: Differentially Private Gradient Ascent-based Mean Shift

Input : Data $S = \{X_i\}_{i=1}^n \subset \mathbb{R}^d$, privacy parameters (ε, δ), minibatch size m, steps $T = \lceil \log n \rceil$, bandwidth h, initialization fraction p_0, stepsize $\eta > 0$, gradient clip C_*

Output: Private mode set $\mathcal{M}$

Sample $k = \max(1, \lfloor np_0 \rfloor)$ points from S and denote them as $\mathcal{I} = \{x_{01}, \ldots, x_{0k}\}$.;
$\mathcal{M} \leftarrow \emptyset$;
Compute σ^2 based on $C_*, \varepsilon, \delta, T$;
Compute kernel matrix $\mathbf{K}$ with entries $K_{ij} = \exp(-\|x_{0i} - x_{0j}\|^2/2h^2)$ for $1 \le i, j \le k$;
Simulate dT correlated Gaussian vectors $z_{j,t} \sim \mathcal{N}(0, \sigma^2\mathbf{K})$ for $1 \le j \le d$, $1 \le t \le T$;
for $l \in [k]$ **do**
    Set $x_0 = x_{0l}$; **for** $t = 0$ **to** $T - 1$ **do**
        Sample minibatch $\mathcal{B}_t \subseteq [n]$, $|\mathcal{B}_t| = m$;
        Compute per-sample gradients $q_i(x_t)$ using KDE weights;
        Clip the norm of each $q_i(x_t)$ to C_*;
        Compute mean gradient $\bar{q}(x)$;
        Compute $z_t = (z_{1,t,l},\ z_{2,t,l},\ \ldots,\ z_{d,t,l}) \in \mathbb{R}^d$;
        Compute $x_{t+1} = x_t + \eta(\bar{q}(x_t) + z_t)$;
    Add x_T to $\mathcal{M}$;
Merge nearby outputs in $\mathcal{M}$;
return $\mathcal{M}$

Algorithm 2: DP-PMS: Differentially Private Partial Mean-Shift

Input : Predictors $X = (X_1, \ldots, X_n)$, responses $Y = (Y_1, \ldots, Y_n)$, optional mesh mesh, privacy (ε, δ), iterations T

Output: Private mode estimates $\widehat{Y} = (\widehat{Y}_1, \ldots, \widehat{Y}_k)$

if *mesh not provided* **then**
    set mesh $\leftarrow$ Y
Let $k \leftarrow |\text{mesh}|$;
Compute kernel bandwidth $h \leftarrow \text{ComputeBandwidth}(\text{mesh}, Y)$;
Initialize $\widehat{Y} \leftarrow$ mesh;
Precompute mesh covariance $K_{\text{mesh}} \leftarrow \text{KernelCovariance}(\text{mesh}; h)$;
Choose noise scale $\sigma \leftarrow \text{CalibrateNoise}(n, \varepsilon, \delta, T, h)$;
for $t \leftarrow 1$ **to** T **do**
    for *each mesh index* $i = 1, \ldots, k$ **do**
        $\widehat{Y}_i \leftarrow \text{WeightedAverage}_j(Y_j;\ w_j = \text{Kernel}((X_j, \widehat{Y}_j), (X_i, \widehat{Y}_i); h))$;
    $\widehat{Y} \leftarrow \widehat{Y} + \sigma \cdot Z$, where $Z \sim \mathcal{N}(0,\ K_{\text{mesh}})$;
return $\widehat{Y}$

- **Algorithm 1 (DP-GRAMS)**, already presented in the main paper, implements a gradient ascent-based mean-shift procedure to estimate modes under (ε, δ)-differential privacy.
- **Algorithm 2 (DP-PMS)** extends DP-GRAMS to the modal regression setting. For each point in a predictor mesh, local responses are identified, mean-shift updates are computed

Algorithm 3: DP-GRAMS-C: Differentially Private Clustering via Private Modes

Input : Data $S = \{X_i\}_{i=1}^n \subset \mathbb{R}^d$, privacy parameters $\varepsilon_{\text{modes}}, \delta$, optional number of clusters k_{est}, initialization fraction p_0.

Output: Private cluster centers $\mathcal{M}$, private labels y

Compute DP candidate modes: $\mathcal{M}_{\text{raw}} \leftarrow \text{DP-GRAMS}(S, \varepsilon_{\text{modes}}, \delta, p_0)$;

Refine modes using mean-shift: $\mathcal{M}_{\text{refined}} \leftarrow \text{MeanShift}(\mathcal{M}_{\text{raw}})$;

if *k_{est} specified* **then**

  Merge modes into k_{est} clusters using agglomerative clustering;

else

  Merge modes using a distance-threshold criterion based on kernel bandwidth;

Assign each data point to the nearest merged mode:

$$y_i = \arg\min_j \|X_i - \mathcal{M}_j\|_2$$

return $\mathcal{M}, y$

with clipped gradients, and Gaussian noise is added to guarantee privacy. The output is a set of private conditional mode estimates.

- **Algorithm 3 (DP-GRAMS-C)** estimates candidate modes privately using DP-GRAMS, merges them into k clusters if specified, and assigns class labels to each data point based on the nearest private cluster mode. This procedure provides end-to-end differential privacy while leveraging the cluster structure for label assignment.

3 Proofs

Before presenting our theoretical results and their proofs, we present the required assumptions. In the following, for any radius $r_j > 0$, we write the closed Euclidean ball as

$$\overline{B}(\mu_j, r_j) := \{x \in \mathbb{R}^d : \|x - \mu_j\| \le r_j\}.$$

Assumption 1 (Model assumptions)**.** *Let $d \ge 1$ and let $p : \mathbb{R}^d \to [0, \infty)$ be a probability density. Fix $k \in \mathbb{N}$. Assume that $p(\cdot)$ has k modes at $\mu_1, \dots, \mu_k$.*

For each $j \in [k]$, there exists $r_j > 0$, such that p is strictly positive on the closed ball $\overline{B}(\mu_j, r_j)$, with

$$p_{\min,j} := \inf_{x \in \overline{B}(\mu_j, r_j)} p(x) > 0.$$

Moreover, $p \in C^3(\overline{B}(\mu_j, r_j))$ and there exists a finite constant $M_j > 0$ such that

$$\sup_{x \in \overline{B}(\mu_j, r_j)} \max_{|\alpha| \le 3} \|D^\alpha p(x)\| \le M_j,$$

for D^α as the partial derivative of multi-index α.

Assumption 2 (Bandwidth condition)**.** *Let $(h_n)_{n \ge 1}$ be the sequence of bandwidths. We assume*

$$h_n \downarrow 0 \qquad \text{and} \qquad n\, h_n^{d+4} \to \infty \qquad \text{as } n \to \infty.$$

Lemma 1 (Per-Sample Sensitivity Bound)**.** *Under Assumptions 1 and 2, for each $i \in [n]$, the per-sample gradient $q_i(x) = h^2 g_i(x)/\hat{p}_h(x)$ satisfies*

$$\|q_i(x)\| \lesssim \frac{1}{p(x) V_d} h^{1-d}, \quad V_d = \frac{\pi^{d/2}}{\Gamma(1 + d/2)} \tag{1}$$

for every $x \in \overline{B}(\mu_j, r_j)$, with high probability.

The proof of the above lemma is given in Section 3.1.

We recall that at the t-th step of DP-GRAMS, we add noise generated as $z_t \sim \mathcal{N}(0, \sigma^2 I_d)$, where

$$\sigma = \frac{C_*/m}{\log(1 + n(e^{\varepsilon_{\text{iter}}} - 1)/m)} \sqrt{8 \log\left(\frac{2.5mT}{n\delta}\right)}, \tag{2}$$

where

$$C_* = \frac{1}{p_{\min} V_d} h^{1-d} \quad \text{and} \quad \varepsilon_{\text{iter}} = \frac{\varepsilon}{\sqrt{8T \log(2/\delta)}}. \tag{3}$$

When $\varepsilon \ll 1$, one can replace $e^{\varepsilon_{\text{iter}}} - 1$ by $\varepsilon_{\text{iter}}$ and $\log(1 + n\varepsilon_{\text{iter}}/m)$ by $n\varepsilon_{\text{iter}}/m$, which yields

$$\sigma \approx \frac{8\, C_*}{n\, \varepsilon} \sqrt{T \log\left(\frac{2}{\delta}\right) \log\left(\frac{2.5\, m\, T}{n\, \delta}\right)}. \tag{4}$$

Theorem 3.1 (Privacy Guarantee of DP-GRAMS)**.** *Suppose Assumptions 1 and 2 hold. Let x_T be the T-th iterate of Algorithm 1 initialized at $x_0 \in \cup_j \overline{B}(\mu_j, r_j)$, with added noise $z_t \sim \mathcal{N}(0, \sigma^2 I_d)$ at each step, for $1 \le t \le T$ and σ as defined in (2). Then x_T satisfies (ε, δ)-differential privacy.*

Proof of Theorem 3.1. To obtain (ε, δ) for the entire T-step phase we apply the (simplified) advanced composition allocation commonly used in DP-SGD analyses. With the usual slack choice $\delta' = \frac{1}{2}\delta_g$, we set the per-iteration privacy (for the *subsampled* mechanism) to

$$\varepsilon_{\text{iter}} = \frac{\varepsilon_g}{2\sqrt{2T \ln(2/\delta_g)}}, \qquad \delta_{\text{iter}} = \frac{\delta_g}{2T}.$$

At iteration t, a mini-batch $\mathcal{B}_t$ of size m is sampled uniformly without replacement from a dataset of size n. Denote the sampling ratio by $q = \frac{m}{n}$. Per-sample gradients are clipped at norm C_* (cf. (3)), so the ℓ_2-sensitivity of the average clipped gradient on a batch of size m satisfies

$$\Delta_g \le \frac{2C_*}{m}.$$

By the amplification theorem for sampling without replacement (see Balle et al. (2018)), if the full-data per-iteration mechanism is $(\varepsilon^*, \delta^*)$-DP, then the subsampled mechanism has privacy budget

$$\varepsilon_{\text{iter}} = \ln(1 + q(e^{\varepsilon^*} - 1)), \qquad \delta_{\text{iter}} = q\, \delta^*.$$

These relations are inverted exactly to give the full-data per-iteration privacy budget:

$$\varepsilon^* = \ln\Big(1 + \frac{e^{\varepsilon_{\text{iter}}} - 1}{q}\Big), \qquad \delta^* = \frac{\delta_{\text{iter}}}{q} = \frac{\delta}{2Tq} = \frac{n\delta}{2mT}.$$

By the Gaussian mechanism ((Dwork et al., 2014, Theorem 3.22)), to achieve $(\varepsilon^*, \delta^*)$-DP for a query of ℓ_2-sensitivity Δ, one must add Gaussian noise $z \sim \mathcal{N}(0, \sigma^2 I)$ with standard deviation

$$\sigma = \frac{\Delta}{\varepsilon^*} \sqrt{2 \ln\Big(\frac{1.25}{\delta^*}\Big)} = \frac{2C_*/m}{\ln\Big(1 + \frac{e^{\varepsilon_{\text{iter}}} - 1}{q}\Big)} \sqrt{2 \ln\left(\frac{1.25}{\frac{n\delta}{2mT}}\right)}.$$

Therefore, the exact calibration of σ expressed directly in terms of the total budget (ε, δ) is

$$\sigma = \frac{\frac{2C_*}{m}}{\ln\left(1 + \frac{1}{q}\left(\exp\left(\frac{\varepsilon}{2\sqrt{2T \ln(2/\delta)}}\right) - 1\right)\right)} \sqrt{2 \ln\left(\frac{2.5mT}{n\delta}\right)}. \tag{5}$$

This completes the proof. □

Remark 3.1 (Correlated Noise.). *The use of multiple initializations implies that the same data is used several times throughout the algorithm, leading to privacy leakage at the post-processing agglomeration stage detailed below. To avoid this, at each step of gradient descent we use noise vectors that are independent across iterations and data dimensions, but correlated across initializations. More specifically, we think of the iterates* $(x_0, x_1, \ldots, x_T)$ *as a function of* x_0 *and add suitably correlated noise for two initializations that are close to each other. The exact value of correlation at two initializations* x_0 *and* x_0' *is taken to be* $\exp(-\|x_0 - x_0'\|^2/2h^2)$*, based on the Gaussian kernel. This choice is based on the framework introduced for private kernel density estimation in Hall et al. (2013).*

We now add a brief justification for the above remark. We treat the sequence of iterates $(x_1, \ldots, x_T)$ as a function $f : \mathbb{R}^d \mapsto \mathbb{R}^{d\times T}$, where x_0 is any initialization and $f(x_0|\mathcal{X}) := (x_1, \ldots, x_T)$. Now we can use the technique for privatizing functions as given in Proposition 8 and Corollary 9 of Hall et al. (2013). In particular, since each iterate is a kernel average divided by an $O(1)$ quantity, we can bound the sensitivity $\|f(x_0|\mathcal{X}) - f(x_0|\mathcal{X}')\|$ using the RKHS norm induced by the kernel $K(\cdot)$, by the techniques of Hall et al. (2013), through successive differences among the iterates. This allows us to use the results from Hall et al. (2013) from where the privacy guarantee using correlated noise follows.

We now move on to determine the rate of estimation of the modes through DP-GRAMS. Our results rely on the fact that

$$\ell(x) = \log p(x) \quad \text{and} \quad \hat{\ell}_h(x) = \log \hat{p}_h(x)$$

and the respective first two derivatives would be close. More specifically, let $\mathcal{G}_n^{\text{local},j}$ be the event that

$$\sup_{x\in\overline{B}(\mu_j,r_j)} |\hat{\ell}_h^{(k)}(x) - \ell^{(k)}(x)| < C\left[h^{(2-k)\wedge 2} + \sqrt{\frac{\log n}{nh^{d+2k}}}\right] \tag{6}$$

for some constant $C > 0$ and $k = 0, 1, 2$. Here $\hat{\ell}_h^{(k)}$, $\ell^{(k)}$ denote the k-th derivatives of $\hat{\ell}_h$ and ℓ respectively.

Assumption 3 (Hessian conditions).

a. For each $j \in [k]$,

$$\alpha_j := -\lambda_{\max}(\nabla^2 \log p(\mu_j)) > 0.$$

b. For each $j \in [k]$*, there is a constant* $L_j > 0$ *such that*

$$\sup_{x,y\in\overline{B}(\mu_j,r_j)} \|\nabla^2 \log p(x) - \nabla^2 \log p(y)\| \le L_j\|x - y\|.$$

Assumption 4 (Radius constraint). *For each* $j \in [k]$*, Assumption 1 is satisfied for some*

$$r_j \le \min\left\{\frac{\alpha_j}{2L_j},\ \frac{1}{2}\min_{i\neq j}\|\mu_i - \mu_j\|\right\}.$$

Lemma 2 (Local Hessian bound for $\log p$). *Under Assumption 1, for every* $j \in [k]$ *and every* $x \in \overline{B}(\mu_j, r_j)$*,*

$$\nabla^2 \log p(x) \preceq -(\alpha_j - L_j r_j)I. \tag{7}$$

In particular, if $r_j \le \dfrac{\alpha_j}{2L_j}$*, then for all* $x \in \overline{B}(\mu_j, r_j)$

$$\nabla^2 \log p(x) \preceq -\tfrac{\alpha_j}{2}I,$$

i.e. $\log p$ *is* $\alpha_j/2$*-strongly concave on* $\overline{B}(\mu_j, r_j)$*.*

Lemma 3 (Mode exclusion). *If*

$$r_j \le \tfrac{1}{2} \min_{i \neq j} \Delta_i^{(j)},$$

then the closed ball $\bar{B}(\mu_j, r_j) = \{x : \|x - \mu_j\| \le r_j\}$ *contains no other mode* μ_i *with* $i \neq j$.

Lemma 4 (Uniform-in-x log-KDE rates on local balls). *Let* $X_1, \dots, X_n$ *be i.i.d. with density* p *satisfying Assumption 1 and Assumption (2). Let* $K(u) = (2\pi)^{-d/2} \exp(-\|u\|^2/2)$ *be the Gaussian kernel, and let Assumption 2 hold true. Fix a mode index* $j \in [k]$ *and the closed ball* $\overline{B}(\mu_j, r_j)$ *satisfying Assumption 4. Then*

$$\sup_{x \in \overline{B}(\mu_j, r_j)} |\hat{\ell}_h(x) - \log p(x)| = O(h^2) + O\Big(\sqrt{\frac{\log n}{n h^d}}\Big),$$

$$\sup_{x \in \overline{B}(\mu_j, r_j)} \|\nabla \hat{\ell}_h(x) - \nabla \log p(x)\| = O(h^2) + O\Big(\sqrt{\frac{\log n}{n h^{d+2}}}\Big),$$

$$\sup_{x \in \overline{B}(\mu_j, r_j)} \|\nabla^2 \hat{\ell}_h(x) - \nabla^2 \log p(x)\| = O(h^2) + O\Big(\sqrt{\frac{\log n}{n h^{d+4}}}\Big),$$

with probability at least $1 - n^{-4}$.

Lemma 5 (Local KDE lower bound). *Let* $X_1, \dots, X_n \overset{i.i.d.}{\sim} p$ *satisfy Assumption 1, and fix* $j \in [k]$ *with corresponding radius* $r_j > 0$. *Let the Gaussian kernel density estimator with bandwidth* $h > 0$ *be*

$$\hat{p}_h^*(x) := \frac{1}{n} \sum_{i=1}^{n} \exp\Big(-\frac{\|X_i - x\|^2}{2h^2}\Big), \qquad x \in \overline{B}(\mu_j, r_j).$$

Then, there exists a constant $c_j > 0$ *such that, for all sufficiently large* n,

$$\Pr\Big(\hat{p}_h^*(x) \ge c_j\, h^d \ \ \forall x \in \overline{B}(\mu_j, r_j)\Big) \ge 1 - 1/n^2.$$

Definition 3.2 (Local good event). *Fix a mode index* $j \in [k]$, *a bandwidth* $h > 0$, *and* $\delta \in (0, 1)$. *We define the* local good event $\mathcal{G}_n^{\mathrm{local},j}$ *to be the event that the following hold uniformly on* $\overline{B}(\mu_j, r_j)$*:*

$$\sup_{x \in \overline{B}(\mu_j, r_j)} |\hat{\ell}_h(x) - \log p(x)| \le C\Big(h^2 + \sqrt{\tfrac{\log n}{n h^d}}\Big),$$

$$\sup_{x \in \overline{B}(\mu_j, r_j)} \|\nabla \hat{\ell}_h(x) - \nabla \log p(x)\| \le C\Big(h^2 + \sqrt{\tfrac{\log n}{n h^{d+2}}}\Big),$$

$$\sup_{x \in \overline{B}(\mu_j, r_j)} \|\nabla^2 \hat{\ell}_h(x) - \nabla^2 \log p(x)\| \le C\Big(h^2 + \sqrt{\tfrac{\log n}{n h^{d+4}}}\Big),$$

and

$$\hat{p}_h^*(x) \ge c\, h^d \quad \forall x \in \overline{B}(\mu_j, r_j),$$

for some constants $C, c > 0$ *(depending on* j *but not on* n*).*

Lemma 6 (High-probability local good event). *Under Assumptions 1 and 2, for each* $j \in [k]$ *there exists* n_0 *such that for all* $n \ge n_0$,

$$\Pr\left(\mathcal{G}_n^{\mathrm{local},j}\right) \ \ge \ 1 - n^{-4}.$$

Theorem 3.3 (Local convergence of DP-GRAMS)**.** *Suppose Assumptions 1, 2, 3, 4 hold, and the initialization satisfies $x_0 \in \overline{B}(\mu_j, r_j)$. Then the output x_T after T steps of DP-GRAMS satisfies:*

$$\mathbb{E}\left[\|x_T - \mu_j\|^2 | \mathcal{X}\right] \le C\Big(\frac{\log n}{n}\Big)^{4/(d+6)} + \Big(\frac{\mathrm{polylog}(n,\delta)}{n^2\varepsilon^2}\Big)^{2/(d+3)}$$

for $T = C\log n$ and some numerical constant $C > 0$, provided $\mathcal{X} \in \mathcal{G}_n^{\mathrm{local},j}$, where $\mathcal{G}_n^{\mathrm{local},j}$ is defined in (6).

Proof of Theorem 3.3. Condition on the observed sample $\mathcal{X} = \{X_1, \dots, X_n\}$ and on the local good event $\mathcal{G}_n^{\mathrm{local}}$ (Definition 3.2) throughout. Let an initialization x_0 lie in the basin of attraction of mode μ_j for some $j \in [k]$, i.e. $x_0 \in \overline{B}(\mu_j, r_j)$.

Write $\delta_t := x_t - \mu_j$.

At iteration t, draw a minibatch $\mathcal{B}_t \subset \mathcal{X}$ of size m uniformly at random without replacement and set

$$\bar{q}_t = \frac{1}{m}\sum_{i\in\mathcal{B}_t} h^2 \frac{g_i(x_t)}{\hat{p}_h(x_t)} = \frac{1}{m}\sum_{i\in\mathcal{B}_t} q_i(x_t), \qquad q_i(x_t) = h^2\frac{g_i(x_t)}{\hat{p}_h(x_t)}.$$

Define

$$q_i^{\mathrm{clip}}(x_t) = q_i(x_t)\mathbf{1}\{\|q_i(x_t)\| \le C_*\} + C_*\frac{q_i(x_t)}{\|q_i(x_t)\|}\mathbf{1}\{\|q_i(x_t)\| > C_*\}, \quad i \in [n],$$

and the per-sample clipping bias

$$b_i(x_t) = q_i^{\mathrm{clip}}(x_t) - q_i(x_t) = \Big(\frac{C_*}{\|q_i(x_t)\|} - 1\Big)q_i(x_t)\mathbf{1}\{\|q_i(x_t)\| > C_*\}.$$

Then

$$\bar{q}_t^{\mathrm{clip}} = \bar{q}_t + \bar{b}_t, \qquad \bar{b}_t = \frac{1}{m}\sum_{i\in\mathcal{B}_t} b_i(x_t).$$

The DP noise is $z_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2 I_d)$, independent of $\bar{q}_t$ across t, and the update is

$$x_{t+1} = x_t + \eta(\bar{q}_t + \bar{b}_t + z_t).$$

From the update,

$$\begin{aligned}\|\delta_{t+1}\|^2 = \|\delta_t\|^2 + \eta^2\|\bar{q}_t\|^2 + \eta^2\|z_t\|^2 + 2\eta\langle\delta_t, \bar{q}_t\rangle + 2\eta\langle\delta_t, z_t\rangle + 2\eta^2\langle\bar{q}_t, z_t\rangle \\ + \eta^2\|\bar{b}_t\|^2 + 2\eta^2\langle\bar{b}_t, z_t\rangle + 2\eta^2\langle\bar{q}_t, \bar{b}_t\rangle + 2\eta\langle\delta_t, \bar{b}_t\rangle.\end{aligned}$$

Taking conditional expectation over the minibatch and z_t (holding $x_t, \mathcal{X}$ fixed), using $\mathbb{E}[z_t] = 0$, $\mathbb{E}|z_t|^2 = d\sigma^2$ and independence of $\bar{q}_t$ and z_t, yields

$$\begin{aligned}&\mathbb{E}[\|\delta_{t+1}\|^2 \mid x_t, \mathcal{X}, \mathcal{G}_n^{\mathrm{local}}] \\ &= \|\delta_t\|^2 + \eta^2\mathbb{E}[\|\bar{q}_t\|^2 \mid x_t, \mathcal{X}, \mathcal{G}_n^{\mathrm{local}}] + \eta^2 d\sigma^2 + 2\eta\langle\delta_t, \mathbb{E}[\bar{q}_t \mid x_t, \mathcal{X}, \mathcal{G}_n^{\mathrm{local}}]\rangle \\ &\quad + \eta^2\mathbb{E}[\|\bar{b}_t\|^2 \mid x_t, \mathcal{X}, \mathcal{G}_n^{\mathrm{local}}] + 2\eta^2\mathbb{E}[\langle\bar{q}_t, \bar{b}_t\rangle \mid x_t, \mathcal{X}, \mathcal{G}_n^{\mathrm{local}}] \\ &\quad + 2\eta\langle\delta_t, \mathbb{E}[\bar{b}_t \mid x_t, \mathcal{X}, \mathcal{G}_n^{\mathrm{local}}]\rangle \\ &\le \|\delta_t\|^2 + \eta^2 T_1 + \eta^2 d\sigma^2 + 2\eta T_2 + \eta^2 T_3 + 2\eta^2\sqrt{T_1 T_3} + 2\eta\|\delta_t\|\sqrt{T_3} \\ &\le \|\delta_t\|^2 + 2\eta^2 T_1 + \eta^2 d\sigma^2 + 2\eta T_2 + 2\eta^2 T_3 + 2\eta\|\delta_t\|\sqrt{T_3} \\ &\le \|\delta_t\|^2 + 2\eta^2 T_1 + \eta^2 d\sigma^2 + 2\eta T_2 + 2\eta^2 T_3 + 2\eta\left(\xi\|\delta_t\|^2 + \frac{1}{4\xi}T_3\right), \end{aligned} \tag{8}$$

where $\xi = h^2\alpha_j/2$ and

$$T_1 := \mathbb{E}[\|\bar{q}_t\|^2 \mid x_t, \mathcal{X}, \mathcal{G}_n^{\mathrm{local}}], \quad T_2 := \langle\delta_t, \mathbb{E}[\bar{q}_t \mid x_t, \mathcal{X}, \mathcal{G}_n^{\mathrm{local}}]\rangle, \quad \text{and} \quad T_3 := \mathbb{E}[\|\bar{b}_t\|^2 \mid x_t, \mathcal{X}, \mathcal{G}_n^{\mathrm{local}}]. \tag{9}$$

We now bound each of these terms one by one using the following lemmas.

Lemma 7.

$$T_1 := \mathbb{E}[\|\bar{q}_t\|^2 \mid x_t, \mathcal{X}, \mathcal{G}_n^{\text{local}}]$$
$$\lesssim \frac{n-m}{m(n-1)} O_p(h^{2-d}) + \Big(1 - \frac{n-m}{m(n-1)}\Big)\Big(h^4\alpha_j^2\|\delta_t\|^2 + O_p\Big(h^8 + h^4\tfrac{\log n}{nh^{d+4}}\Big)\|\delta_t\|^2 + O_p\Big(h^8 + h^4\tfrac{\log n}{nh^{d+2}}\Big)\Big)$$

Lemma 8. $T_3 := \mathbb{E}[\|\bar{b}_t\|^2 \mid x_t, \mathcal{X}, \mathcal{G}_n^{\text{local}}] = O_p(h^9)$.

Lemma 9. *For $x_t \in \overline{B}(\mu_j, r_j)$, $k_n(x) := h^2\nabla \log \hat{p}_h(x)$ satisfies:*

1. $k_n(x_t) = h^2\Big(\int_0^1 \nabla^2 \log \hat{p}_h(\mu_j + s\delta_t)\, ds\Big)\delta_t + O_p\Big(h^4 + h^2\sqrt{\tfrac{\log n}{nh^{d+2}}}\Big)$

2. $\|k_n(x_t)\|^2 \le\ 2h^4\alpha_j^2\|\delta_t\|^2 + O_p\Big(h^8 + h^4\tfrac{\log n}{nh^{d+4}}\Big)\|\delta_t\|^2 + O_p\Big(h^8 + h^4\tfrac{\log n}{nh^{d+2}}\Big)$.

Finally for T_2 using Lemma 9,

$$\begin{aligned}
\langle \delta_t, \mathbb{E}\bar{q}_t\rangle &= \langle \delta_t, k_n(x_t)\rangle \\
&= \langle \delta_t, h^2\Big(\int_0^1 \nabla^2 \log \hat{p}_h(\mu_j + s\delta_t)\, ds\Big)\delta_t + O_p\Big(h^4 + h^2\sqrt{\tfrac{\log n}{nh^{d+2}}}\Big)\rangle \\
&= \langle \delta_t, O_p\Big(h^4 + h^2\sqrt{\tfrac{\log n}{nh^{d+2}}}\Big)\rangle + \langle \delta_t, h^2\Big(\int_0^1 \nabla^2 \log \hat{p}_h(\mu_j + s\delta_t)\, ds\Big)\delta_t\rangle \\
&\le h^2\left(O_p\Big(h^2 + \sqrt{\tfrac{\log n}{nh^{d+2}}}\Big)\|\delta_t\|\right) + h^2\delta_t^\top\Big(\int_0^1 \nabla^2 \log \hat{p}_h(\mu_j + s\delta_t)\, ds\Big)\delta_t \\
&= h^2\left(O_p\Big(h^2 + \sqrt{\tfrac{\log n}{nh^{d+2}}}\Big)\|\delta_t\|\right) \\
&+ h^2\delta_t^\top\Big(\int_0^1 \nabla^2 \log \hat{p}_h(\mu_j + s\delta_t) - \nabla^2 \log p(\mu_j + s\delta_t) + \nabla^2 \log p(\mu_j + s\delta_t)\, ds\Big)\delta_t \\
&\le h^2\left(O_p\Big(h^2 + \sqrt{\tfrac{\log n}{nh^{d+2}}}\Big)\|\delta_t\|\right) + h^2 O_p\Big(h^2 + \sqrt{\tfrac{\log n}{nh^{d+4}}}\Big)\|\delta_t\|^2 - h^2(\frac{\alpha_j}{2})\|\delta_t\|^2 \\
&\le h^2\left(\frac{\alpha_j}{8}\|\delta_t\|^2 + \frac{2}{\alpha_j}O_p\left(h^4 + \frac{\log n}{nh^{d+2}}\right)\right) \\
&+ h^2 O_p\Big(h^2 + \sqrt{\tfrac{\log n}{nh^{d+4}}}\Big)\|\delta_t\|^2 - h^2(\frac{\alpha_j}{2})\|\delta_t\|^2 \\
&\le h^2\left(-\tfrac{3\alpha_j}{8}\|\delta_t\|^2 + \tfrac{2}{\alpha_j}O_p\Big(h^4 + \tfrac{\log n}{nh^{d+2}}\Big)\right) + h^2 O_p\Big(h^2 + \sqrt{\tfrac{\log n}{nh^{d+4}}}\Big)\|\delta_t\|^2,
\end{aligned}$$

on $\mathcal{G}_n^{\text{local}}$.

Inserting the drift and minibatch second-moment bounds (and absorbing the exponentially small clipping contributions into the negligible $O_p(\cdot)$ remainders by the bounds above) into (8) yields (on $\mathcal{G}_n^{\text{local}}$)

$$\begin{aligned}
\mathbb{E}[\|\delta_{t+1}\|^2 \mid x_t, \mathcal{X}, \mathcal{G}_n^{\text{local}}] \lesssim\ & \|\delta_t\|^2 + \eta^2\frac{n-m}{m(n-1)}O_p(h^{2-d}) \\
&+ \eta^2\Big(1 - \frac{n-m}{m(n-1)}\Big)\Big[\tfrac{3}{4}h^4\alpha_j^2\|\delta_t\|^2 + O_p\Big(h^8 + h^4\tfrac{\log n}{nh^{d+4}}\Big)\|\delta_t\|^2\Big] \\
&+ \eta^2 d\sigma^2 \\
&+ 2\eta\Big[-\tfrac{3\alpha_j}{8}h^2\|\delta_t\|^2 + O_p\Big(h^4 + h^2\sqrt{\tfrac{\log n}{nh^{d+4}}}\Big)\|\delta_t\|^2 + \tfrac{2}{\alpha_j}O_p\Big(h^6 + h^2\tfrac{\log n}{nh^{d+2}}\Big)\Big],
\end{aligned} \tag{10}$$

on $\mathcal{G}_n^{\text{local}}$. Collecting terms proportional to $\|\delta_t\|^2$ and constant terms yields the affine inequality (on $\mathcal{G}_n^{\text{local}}$)

$$\mathbb{E}[\|\delta_{t+1}\|^2 \mid x_t, \mathcal{X}, \mathcal{G}_n^{\text{local}}] \le a_n\|\delta_t\|^2 + b_n,$$

with

$$\begin{aligned} a_n &= 1 + \eta^2\Big(1 - \frac{n-m}{m(n-1)}\Big)\Big(h^4\tfrac{3\alpha_j^2}{4} + O_p\Big(h^8 + h^4\tfrac{\log n}{nh^{d+4}}\Big)\Big) + 2\eta\xi \\ &\quad + 2\eta\Big(-\tfrac{3\alpha_j}{8}h^2 + O_p\Big(h^4 + h^2\sqrt{\tfrac{\log n}{nh^{d+4}}}\Big)\Big) \\ &= 1 - \eta h^2(\tfrac{\alpha_j}{4} - R_n) + \mathcal{O}(\eta^2 h^4), \qquad R_n := O_p\Big(h^2 + \sqrt{\tfrac{\log n}{nh^{d+4}}}\Big), \end{aligned} \tag{11}$$

and

$$\begin{aligned} b_n &= \eta^2\frac{n-m}{m(n-1)}O_p(h^{2-d}) + \eta^2\Big(1 - \frac{n-m}{m(n-1)}\Big)O_p\Big(h^8 + h^4\tfrac{\log n}{nh^{d+2}}\Big) \\ &\quad + \eta^2 d\sigma^2 + \tfrac{4\eta}{\alpha_j}O_p\Big(h^6 + h^2\tfrac{\log n}{nh^{d+2}}\Big) + \eta^2 T_3 + \frac{\eta}{4\xi}T_3 \\ &= \eta^2\frac{n-m}{m(n-1)}O_p(h^{2-d}) + \eta^2 d\sigma^2 + \Big(\tfrac{4\eta}{\alpha_j}h^2 + \eta^2 h^4\Big)O_p\Big(h^4 + \tfrac{\log n}{nh^{d+2}}\Big) + 2\eta^2 T_3 + \frac{\eta}{4\xi}T_3 \\ &\le \eta^2\frac{n-m}{m(n-1)}O_p(h^{2-d}) + \eta^2 d\sigma^2 + \Big(\tfrac{4\eta}{\alpha_j}h^2 + \eta^2 h^4\Big)O_p\Big(h^4 + \tfrac{\log n}{nh^{d+2}}\Big) + 2\eta^2 h^9 + \frac{\eta}{2h^2\alpha_j}h^9 \\ &= \eta^2\frac{n-m}{m(n-1)}O_p(h^{2-d}) + \eta^2 d\sigma^2 + \Big(\tfrac{4\eta}{\alpha_j}h^2 + \eta^2 h^4\Big)O_p\Big(h^4 + \tfrac{\log n}{nh^{d+2}}\Big). \end{aligned} \tag{12}$$

all equalities/inequalities holding on $\mathcal{G}_n^{\text{local}}$. (Any explicit clipping contributions are exponentially small in nh^d and are absorbed into the $O_p(\cdot)$ terms above.)

Taking total expectation over x_t (conditioning on $\mathcal{X}$ and $\mathcal{G}_n^{\text{local}}$) and writing $\Delta_t := \mathbb{E}[|\delta_t|^2 \mid \mathcal{X}, \mathcal{G}_n^{\text{local}}]$ yields the scalar recurrence

$$\Delta_{t+1} \le a_n\Delta_t + b_n \quad \text{on } \mathcal{G}_n^{\text{local}}, \tag{13}$$

with solution for any $T \ge 1$:

$$\Delta_T \le a_n^T\Delta_0 + \frac{b_n(1-a_n^T)}{1-a_n} \quad \text{on } \mathcal{G}_n^{\text{local}}. \tag{14}$$

To analyze (14) we require $0 < a_n < 1$. Expanding a_n gives

$$a_n = 1 - \eta h^2\Big(\tfrac{\alpha_j}{4} - R_n\Big) + \mathcal{O}(\eta^2 h^4),$$

where $R_n = O_p\Big(h^2 + \sqrt{\tfrac{\log n}{nh^{d+4}}}\Big) = o_p(1)$ under Assumption 2. Thus, w.h.p. (on $\mathcal{G}_n^{\text{local}}$) the leading linear term is negative. For $a_n > 0$ it suffices that

$$1 - \eta h^2\Big(\tfrac{\alpha_j}{4} - R_n\Big) + \mathcal{O}(\eta^2 h^4) > 0,$$

which (for η in the regime of interest and using Assumption 2) is implied by $\eta \lesssim 1/h^2$. Hence, there exists a constant (depending on α_j and the $O_p(\cdot)$-terms) such that

$$0 < \eta \le C/h^2$$

ensures

$$0 < a_n < 1 \quad \text{w.h.p.}$$

Under this scaling $a_n^T \approx \exp(-c\eta h^2 T)$ with $c = \frac{\alpha_j}{4} - R_n > 0$, so the transient term $a_n^T\Delta_0$ vanishes provided $\eta h^2 T \to \infty$. A convenient choice is $T \asymp \log n$, which yields $a_n^T \to 0$ on $\mathcal{G}_n^{\text{local}}$. Therefore, under these choices,

$$\Delta_T \lesssim \frac{b_n}{1-a_n} \quad \text{on } \mathcal{G}_n^{\text{local}}.$$

We next simplify $\frac{b_n}{1-a_n}$ asymptotically. Collecting dominant contributions and substituting the standard DP Gaussian calibration in (4)

$$\sigma \asymp \frac{C_*}{n\varepsilon}\sqrt{T\ln(2/\delta)\cdot\ln(2.5mT/(n\delta))},$$

one obtains (on $\mathcal{G}_n^{\text{local}}$)

$$\begin{aligned}\frac{b_n}{1-a_n} \asymp \eta\cdot\frac{O_p(h^{-d})}{m} + \eta h^2\cdot O_p\Big(h^4+\tfrac{\log n}{nh^{d+2}}\Big)\\ + \frac{\eta T d}{n^2\varepsilon^2}\cdot\frac{1}{p_{\min,j}^2 V_d^2}h^{-2d}\,\text{polylog} + O_p\Big(h^4+\tfrac{\log n}{nh^{d+2}}\Big),\end{aligned} \tag{15}$$

where $\text{polylog} = 64\ln(2/\delta)\ln(2.5mT/(n\delta))$ and $V_d = \pi^{d/2}/\Gamma(1+d/2)$.

Using $m=\gamma n$ for some fixed $\gamma\in(0,1]$ and simplifying the negligible terms under Assumption 2 gives the leading asymptotics (on $\mathcal{G}_n^{\text{local}}$)

$$\frac{b_n}{1-a_n} \asymp \frac{\eta T d}{n^2\varepsilon^2}\frac{1}{p_{\min,j}^2 V_d^2}h^{-2d} + O_p\Big(h^4+\tfrac{\log n}{nh^{d+2}}\Big).$$

We now distinguish regimes depending on the privacy parameter ε. The DP-dominated regime is defined by the condition that the DP-noise term dominates the classical KDE variance term, i.e.

$$\frac{Td}{n^2\varepsilon^2}\,h^{-2(d+1)}\cdot\text{polylog} \gtrsim \frac{\log n}{nh^{d+2}}.$$

Solving this inequality for ε yields the (upper) condition for the DP regime:

$$\varepsilon \lesssim \sqrt{\frac{Td\cdot\text{polylog}}{n\log n}}\,h^{-d/2}. \tag{16}$$

Under (16) the KDE variance is negligible and the steady-state bound simplifies (on $\mathcal{G}_n^{\text{local}}$) to

$$\frac{b_n}{1-a_n} \asymp h^4 + \frac{Td\cdot\text{polylog}}{n^2\varepsilon^2}\,h^{-2d}.$$

Balancing the two terms yields the DP-optimal bandwidth and MSE. Writing $K := Td\cdot\text{polylog}$, minimize

$$f(h) := h^4 + \frac{K}{h^{2d+2}n^2\varepsilon^2}.$$

Setting $f'(h)=0$ gives

$$h_{\text{opt}}^{\text{DP}} \asymp \Big(\frac{K}{n^2\varepsilon^2}\Big)^{1/(2d+6)}, \qquad \text{MSE}_{\text{opt}}^{\text{DP}} \asymp \Big(\frac{K}{n^2\varepsilon^2}\Big)^{2/(d+3)},$$

all statements holding on $\mathcal{G}_n^{\text{local}}$.

In contrast, if the DP-noise term is smaller than the KDE variance, we are in the non-DP regime and the dominant contributions are kernel bias and finite-sample variance:

$$\frac{b_n}{1-a_n} \asymp O_p\Big(h^4+\tfrac{\log n}{nh^{d+2}}\Big) \quad \text{on } \mathcal{G}_n^{\text{local}}.$$

Balancing these terms yields the classical non-DP choices

$$h_{\text{opt}}^{\text{non-DP}} \asymp \Big(\tfrac{\log n}{n}\Big)^{1/(d+6)}, \qquad \text{MSE}_{\text{opt}}^{\text{non-DP}} \asymp \Big(\tfrac{\log n}{n}\Big)^{4/(d+6)}.$$

To compare the two regimes and obtain explicit thresholds in ε, evaluate the dominance inequality at the non-DP optimal bandwidth $h_{\text{opt}}^{\text{non-DP}}$. Denote again $K := Td \cdot \text{polylog}$. The inequality

$$\frac{K}{n^2\varepsilon^2}\, h^{-2(d+1)} \gtrsim \frac{\log n}{nh^{d+2}}$$

at $h = h_{\text{opt}}^{\text{non-DP}} \asymp (\log n/n)^{1/(d+6)}$ yields the lower cutoff

$$\varepsilon_{\text{non}} \asymp \frac{\sqrt{K}}{n^{3/(d+6)} \log n}. \tag{17}$$

Equating the DP-optimal MSE $\left(K/(n^2\varepsilon^2)\right)^{2/(d+3)}$ with the non-DP optimal MSE $(\log n/n)^{4/(d+6)}$ yields the (larger) cutoff

$$\varepsilon_{\text{DP}} \asymp \sqrt{K}\, n^{-3/(d+6)} (\log n)^{-(d+3)/(d+6)}. \tag{18}$$

Since $\varepsilon_{\text{DP}}/\varepsilon_{\text{non}} \asymp (\log n)^{3/(d+6)} > 1$, we obtain the trichotomy:

- If $\varepsilon \ll \varepsilon_{\text{non}}$ (strong privacy), the DP-noise term already dominates at the non-DP bandwidth: use $h_{\text{opt}}^{\text{DP}}$ and achieve $\text{MSE}_{\text{opt}}^{\text{DP}} \asymp \left(K/(n^2\varepsilon^2)\right)^{2/(d+3)}$ (on $\mathcal{G}_n^{\text{local}}$).
- If $\varepsilon_{\text{non}} \ll \varepsilon \ll \varepsilon_{\text{DP}}$ (intermediate window), the non-DP bandwidth still balances classical bias/variance but DP noise prevents attaining the non-DP optimal MSE.
- If $\varepsilon \gg \varepsilon_{\text{DP}}$ (weak privacy), DP noise is asymptotically negligible and the classical non-DP bandwidth and MSE govern the rate.

All high-probability statements above are conditional on $\mathcal{G}_n^{\text{local}}$; by Definition 3.2 and Lemma 4 there exists n_0 such that for all $n \geq n_0$ the event $\mathcal{G}_n^{\text{local}}$ holds with probability at least $1 - n^{-2}$, so the conclusions hold with the same probability. This completes the proof. □

Let us define the in-ball probabilities

$$p_j := \Pr_{X \sim p}\left(X \in \bar{B}(\mu_j, r_j)\right) \quad \text{for} \quad j \in [k].$$

The next proposition shows that with high probability, random initialization provides at least one point in $\bar{B}(\mu_j, r_j)$ for every $j \in [k]$.

Proposition 3.4 (Initialization covers all modes)**.** *Under Assumptions 1, 2, 3, 4, with probability at least $1 - n^{-2}$, the initialization pool $\mathcal{I}$ is non-empty and contains at least one point inside each ball $\bar{B}(\mu_j, r_j)$, provided*

$$\frac{n}{\log(kn)} \;\geq\; \frac{C}{p_0} \max_{1 \leq j \leq k} \frac{1}{p_j} \tag{19}$$

for some constant $C > 0$.

Proof of Proposition 3.4. For $i = 1, \ldots, n$ let $w_i \sim \text{Bernoulli}(p_0)$ be the independent indicators used to form $\mathcal{I} = \{X_i : w_i = 1\}$. For each fixed mode index j define

$$I_{i,j} := \mathbf{1}\{X_i \in B(\mu_j, r_j) \text{ and } w_i = 1\}, \qquad i = 1, \ldots, n.$$

The variables $I_{1,j}, \ldots, I_{n,j}$ are iid Bernoulli with success probability $\theta_j := p_0 p_j$. Let us define

$$S_j := \sum_{i=1}^{n} I_{i,j},$$

the number of initialization points falling in $B(\mu_j, r_j)$. Then $S_j \sim \text{Binomial}(n, \theta_j)$ and $\mathbb{E}[S_j] = np_0p_j$. Note that

$$\Pr(S_j = 0) = (1 - p_0p_j)^n \le \exp(-np_0p_j)$$

since $1 - x \le \exp(-x)$ for $x \in [0, 1]$. Now for a fixed $\delta > 0$, if $n \ge \dfrac{1}{p_0p_j} \log(\frac{2k}{\delta})$ then $\Pr(S_j = 0) \le \frac{\delta}{2k}$. By the union bound over $j = 1, \dots, k$,

$$\Pr\Big(\exists j \in \{1, \dots, k\} \text{ with } S_j = 0\Big) \le \sum_{j=1}^{k} \Pr(S_j = 0) \le k \cdot \frac{\delta}{2k} = \frac{\delta}{2}.$$

Next control $\Pr(|\mathcal{I}| = 0)$. The total size $|\mathcal{I}| = \sum_{i=1}^{n} w_i$ is $\text{Binomial}(n, p_0)$, hence

$$\Pr(|\mathcal{I}| = 0) = (1 - p_0)^n \le \exp(-np_0).$$

If $n \ge \dfrac{\log(2/\delta)}{p_0}$ then $\Pr(|\mathcal{I}| = 0) \le \frac{\delta}{2}$.

If n satisfies both bounds in (19) then, by the union bound, the probability that either $|\mathcal{I}| = 0$ or some $S_j = 0$ is at most δ. Therefore, with probability at least $1 - \delta$ we have $|\mathcal{I}| \ge 1$ and $S_j \ge 1$ for every j, i.e., C contains at least one point in each ball $B(\mu_j, r_j)$. Choosing $\delta = 1/n^3$ and applying a union bound over $j \in [k]$ finishes the proof. □

Definition 3.5 (Global good event)**.** *For each mode $j \in [k]$, let $\mathcal{G}_n^{\text{local},j}$ be the local good event on $\overline{B}(\mu_j, r_j)$ from Definition 3.2. Let $\mathcal{I}$ be the initialization pool formed as in Proposition 3.4, and define*

$$\mathcal{E}_{\text{init}} := \{\mathcal{I} \text{ contains at least one point in every ball } \overline{B}(\mu_j, r_j),\ j = 1, \dots, k\}.$$

The global good event *is*

$$\mathcal{G}_n^{\text{global}} := \Big(\bigcap_{j=1}^{k} \mathcal{G}_n^{\text{local},j}\Big) \cap \mathcal{E}_{\text{init}}.$$

Lemma 10 (High-probability global good event)**.** *Under Assumptions 1, 2, 3, and 4, there exists n_0 such that for all $n \ge n_0$,*

$$\Pr\left(\mathcal{G}_n^{\text{global}}\right) \ \ge\ 1 - n^{-2}.$$

Theorem 3.6 (Global convergence of DP-GRAMS)**.** *Suppose Assumptions 1, 2, 3, 4 hold. Let $\mathcal{I}$ be the initialization pool from Proposition 3.4. Then, with probability at least $1 - n^{-1}$, for every $j \in [k]$ there exists at least one $x_0 \in \mathcal{I}$ such that x_T, the output of Algorithm 1 initialized at x_0, satisfies:*

$$\mathbb{E}\left[\|x_T - \mu_j\|^2 | \mathcal{X}\right] \le C\Big(\frac{\log n}{n}\Big)^{4/(d+6)} + \Big(\frac{\text{polylog}(n, \delta)}{n^2\varepsilon^2}\Big)^{2/(d+3)}$$

for $T = C \log n$ and some numerical constant $C > 0$.

Proof of Theorem 3.6. Let n_0 be large enough so that the conclusion of Definition 3.5 holds (this is obtained by combining Lemma 4 and Proposition 3.4 and applying the union bound as in Definition 3.5). Work on the event $\mathcal{G}_n^{\text{global}}$, which holds with probability at least $1 - n^{-2}$.

By construction of $\mathcal{G}_n^{\text{global}}$, the initialization pool $\mathcal{I}$ contains at least one data point inside each ball $B(\mu_j, r_j)$, $j = 1, \dots, k$ (this is the event $\mathcal{E}_{\text{init}}$ guaranteed by Proposition 3.4). For each j, pick an initialization $x_0^{(j)} \in \mathcal{I} \cap B(\mu_j, r_j)$ (such a point exists on $\mathcal{G}_n^{\text{global}}$). Because $\mathcal{G}_n^{\text{global}}$ also contains the local good events on every $\overline{B}(\mu_j, r_j)$, the assumptions of the local convergence

result (Theorem 3.3, proved conditional on a local good event) are satisfied simultaneously for each j.

Apply Theorem 3.3 to each initialization $x_0^{(j)}$. Theorem 3.3 (conditional on the corresponding local good event) guarantees that the iterate sequence started at $x_0^{(j)}$ remains in the basin $\overline{B}(\mu_j, r_j)$ and achieves the steady-state mean-squared error bound stated there (the bound depends on $h, \eta, m, n, \varepsilon, \delta$ and the local constants at μ_j). Since all k local good events and the initialization event hold simultaneously on $\mathcal{G}_n^{\text{global}}$, the conclusions of Theorem 3.3 hold simultaneously for every $j = 1, \ldots, k$ on $\mathcal{G}_n^{\text{global}}$.

Therefore, conditional on $\mathcal{G}_n^{\text{global}}$ (hence with probability at least $1 - n^{-2}$), DP-GRAMS recovers every mode in the sense that for each true mode μ_j there exists an initialization in its basin and the corresponding iterate sequence converges (up to the steady-state MSE given by Theorem 3.3) to a neighborhood of μ_j. This completes the proof. □

3.1 Proofs of Lemmas

Proof of Lemma 1. By definition, the per-sample gradient contribution is

$$q_i(x) = h^2 \frac{g_i(x)}{\hat{p}_h(x)} = \frac{(X_i - x)\exp\big(-\|X_i - x\|^2/(2h^2)\big)}{\frac{1}{n}\sum_{j=1}^n \exp\big(-\|X_j - x\|^2/(2h^2)\big)}.$$

For any $r > 0$, define $\mathcal{N}(x, r) := \{j : \|X_j - x\| \le r\}$. Then

$$\begin{aligned}
\hat{p}_h^*(x) &:= \frac{1}{n}\sum_{j=1}^n \exp\big(-\|X_j - x\|^2/(2h^2)\big) \\
&\ge \frac{1}{n}\sum_{j \in B(x,r)} \exp(-\|X_j - x\|^2/(2h^2)) \\
&\ge e^{-r^2/(2h^2)}\frac{\#B(x,r)}{n}. \qquad (20)
\end{aligned}$$

Setting $r = h$ yields

$$\hat{p}_h^*(x) \ge e^{-1/2}\frac{\#\mathcal{N}(x,h)}{n}.$$

Let $r_i := \|X_i - x\|$ and define $\phi(r_i) := r_i \exp(-r_i^2/(2h^2))$. The maximum occurs at $r_i = h$ with value $\phi(h) = he^{-1/2}$. Hence,

$$\|(X_i - x)\exp(-\|X_i - x\|^2/(2h^2))\| \le he^{-1/2}.$$

Now assume $x \in \overline{B}(\mu_j, r_j)$ for some $j \in [k]$. By Assumption 1, $x \in \overline{B}(\mu_j, r_j))$ satisfies $p(x) \ge p_{\min,j} > 0$. Consider

$$\mathbb{P}(\|X - x\| \le h) = \int_{B(x,h)} p(y)\, dy.$$

A Taylor expansion of p around x up to second order with remainder gives

$$p(y) = p(x) + \nabla p(x)^\top (y - x) + \tfrac{1}{2}(y - x)^\top \nabla^2 p(x)(y - x) + R_3(x, y),$$

where, by Assumption 1,

$$|R_3(x, y)| \le \frac{M_j}{6}\|y - x\|^3.$$

Integrating over $y \in \bar{B}(x, h)$,

$$\int_{B(x,h)} p(y)\, dy = p(x)V_d h^d + \tfrac{1}{2}\int_{B(x,h)} (y - x)^\top \nabla^2 p(x)(y - x)dy + O(h^{d+3}).$$

The linear term vanishes by symmetry, the quadratic term is $O(h^{d+2})$, and the remainder is $O(h^{d+3})$. Therefore,

$$\int_{B(x,h)} p(y)\,dy = p(x)V_d h^d + O(h^{d+2}).$$

By the law of large numbers and Assumption 2, which requires $h_n \to 0$ and $nh_n^d \to \infty$, we obtain the uniform-in-probability bound

$$\frac{\#\mathcal{N}(x,h_n)}{n} = p(x)V_d h_n^d + O_p(h_n^{d+2}) \geq 0.99 p(x) V_d h_n^d$$

for sufficiently large n. Combining with (20), we have

$$\|q_i(x)\| \leq \frac{nh_n}{\#\mathcal{N}(x,h_n)} \lesssim \frac{1}{p(x)V_d} h_n^{1-d}, \qquad \text{as } n \to \infty.$$

□

Proof of Lemma 2. Fix j and $x \in \overline{B}(\mu_j, r_j)$. By Assumption 3(b) together with $\|x - \mu_j\| \leq r_j$, we obtain

$$\left\|\nabla^2 \log p(x) - \nabla^2 \log p(\mu_j)\right\|_{\text{op}} \leq L_j \|x - \mu_j\| \leq L_j r_j.$$

Since $\nabla^2 \log p(\mu_j) \preceq -\alpha_j I$ by Assumption 3(a), it follows that

$$\nabla^2 \log p(x) \preceq \nabla^2 \log p(\mu_j) + L_j r_j I \preceq -\alpha_j I + L_j r_j I = -(\alpha_j - L_j r_j) I,$$

which is (7). The rest follows immediately if $r_j \leq \alpha_j/(2L_j)$. □

Proof of Lemma 3. Suppose for contradiction that $\mu_i \in \bar{B}(\mu_j, r_j)$ for some $i \neq j$. Then $\|\mu_i - \mu_j\| \leq r_j \leq \frac{1}{2}\min_{i' \neq j} \Delta_{i'}^{(j)} \leq \frac{1}{2}\|\mu_i - \mu_j\|$, which is impossible. Hence, no other mode lies in $\bar{B}(\mu_j, r_j)$. □

Proof of Lemma 4. See discussion above Theorem 5 in Genovese et al. (2014) and references therein. □

Proof of Lemma 5. The proof follows from Lemma 4 and Assumption 1, in particular the fact that $\min_{x \in \overline{B}(\mu_j, r_j)} p(x) \geq p_{\min} > 0$. We provide a proof sketch for completeness.

Define the population kernel density at x,

$$p_h(x) := \mathbb{E}[\hat{p}_h^*(x)] = \mathbb{E}_X\Big[\exp\Big(-\frac{\|X - x\|^2}{2h^2}\Big)\Big].$$

By a change of variables $u = (X - x)/h$ with $dX = h^d du$, we have

$$p_h(x) = h^d \int_{\mathbb{R}^d} \exp(-\|u\|^2/2)\, p(x + hu)\, du.$$

Since p is continuous and strictly positive on $\overline{B}(\mu_j, r_j)$ by Assumption 1, there exists a constant $p_{\min,j} > 0$ such that $p(y) \geq p_{\min,j}$ for all $y \in \overline{B}(\mu_j, r_j)$. Restricting the integral to $\|u\| \leq 1$, which is contained in the ball of radius r_j/h for small h, we obtain

$$p_h(x) \geq h^d \int_{\|u\| \leq 1} \exp(-\|u\|^2/2)\, p_{\min,j}\, du =: c_j\, h^d.$$

Let $Y_i := \exp(-\|X_i - x\|^2/(2h^2)) \in [0,1]$. By Bernstein's inequality, we obtain

$$\Pr\left(\hat{p}_h^*(x) \leq p_h(x)/2\right) \leq 2\exp(-nc_j^2 h^d) \leq \frac{1}{n^4},$$

where the last inequality follows by Assumption 2. To extend the bound uniformly over $x \in \overline{B}(\mu_j, r_j)$, cover the closed ball by a finite $h/2$-net $\{x_\ell\}_{\ell=1}^{M}$, where $M \lesssim (r_j/h)^d$. Applying the above bound to each x_ℓ and taking a union bound over M points yields

$$\Pr\left(\hat{p}_h^*(x_\ell) \geq c_j\, h^d/2 \text{ for all } \ell = 1, \ldots, M\right) \geq 1 - 1/n^2,$$

for sufficiently large n. Continuity of $\hat{p}_h^*$ in x then ensures

$$\hat{p}_h^*(x) \geq c_j\, h^d \quad \forall x \in \overline{B}(\mu_j, r_j),$$

with probability at least $1 - 1/n^2$, completing the proof. □

Proof of Lemma 6. For $k = 0, 1, 2$, let A_k be the uniform log-KDE bound from Lemma 4, and let B be the lower–bound event from Lemma 5.

By adjusting the constants in Lemma 4 and Lemma 5, we may assume

$$\Pr(A_k^c) \leq \tfrac{1}{4} n^{-4} \quad (k = 0, 1, 2), \qquad \Pr(B^c) \leq \tfrac{1}{4} n^{-4},$$

for all $n \geq n_0$.

By the union bound,

$$\Pr((A_0 \cap A_1 \cap A_2 \cap B)^c) \;\leq\; \sum_{k=0}^{2} \Pr(A_k^c) + \Pr(B^c) \;\leq\; n^{-4}.$$

Thus,

$$\Pr(\mathcal{G}_n^{\text{local},j}) = \Pr(A_0 \cap A_1 \cap A_2 \cap B) \geq 1 - n^{-4}.$$

□

Proof of Lemma 7. Because $\mathcal{B}_t$ is a uniformly random m-subset of $\{1, \ldots, n\}$,

$$\Pr(i \in \mathcal{B}_t) = \frac{m}{n}, \qquad \Pr(i, j \in \mathcal{B}_t) = \frac{m(m-1)}{n(n-1)} \quad (i \neq j).$$

Writing the minibatch mean with indicators,

$$\bar{q}_t = \frac{1}{m} \sum_{i=1}^{n} q_i(x_t) \mathbf{1}\{i \in \mathcal{B}_t\},$$

we obtain

$$\mathbb{E}[\bar{q}_t \mid x_t, \mathcal{X}, \mathcal{G}_n^{\text{local}}] = \frac{1}{m} \sum_{i=1}^{n} q_i(x_t) \Pr(i \in \mathcal{B}_t) = \frac{1}{n} \sum_{i=1}^{n} q_i(x_t) = h^2 \nabla \log \hat{p}_h(x_t) =: k_n(x_t).$$

For the second moment, expanding the squared norm and taking expectation yields

$$\mathbb{E}[\|\bar{q}_t\|^2 \mid x_t, \mathcal{X}, \mathcal{G}_n^{\text{local}}] = \frac{1}{m^2} \left(\frac{m}{n} \sum_{i=1}^{n} \|q_i(x_t)\|^2 + \frac{m(m-1)}{n(n-1)} \sum_{i \neq j} q_i(x_t)^\top q_j(x_t) \right).$$

Using the identity $\sum_{i \neq j} q_i^\top q_j = n^2 |k_n(x_t)|^2 - \sum_{i=1}^{n} |q_i(x_t)|^2$ gives the finite-population form

$$\mathbb{E}[\|\bar{q}_t\|^2 \mid x_t, \mathcal{X}, \mathcal{G}_n^{\text{local}}] = \frac{n-m}{mn(n-1)} \sum_{i=1}^{n} \|q_i(x_t)\|^2 + \left(1 - \frac{n-m}{m(n-1)}\right) \|k_n(x_t)\|^2.$$

On the local good event $\mathcal{G}_n^{\text{local}}$, we have the uniform lower bound $\hat{p}_h^*(x) \gtrsim h^d$ for all $x \in \overline{B}(\mu_j, r_j)$ (Lemma 5) and the uniform log-KDE derivative bounds (Lemma 4). Consider the population conditional expectation (given x_t)

$$\mathbb{E}[\|q(X; x_t)\|^2 \mid x_t] = \frac{1}{\hat{p}_h^*(x_t)^2} \int_{y \in \overline{B}(\mu_j, r_j)} \|y - x_t\|^2 e^{-\|y-x_t\|^2/h^2} p(y)\, dy.$$

Changing variables $u = (y - x_t)/h$ and $dy = h^d du$ gives

$$\begin{aligned}\mathbb{E}[\|q(X; x_t)\|^2 \mid x_t] &= \frac{h^{d+2}}{\hat{p}_h^*(x_t)^2} \int_{x_t + hu \in \overline{B}(\mu_j, r_j)} \|u\|^2 e^{-\|u\|^2} p(x_t + hu)\, du \\ &\lesssim \frac{h^{d+2}}{\hat{p}_h^*(x_t)^2}. \end{aligned} \tag{21}$$

Since $\hat{p}_h^*(x_t) \gtrsim h^d$, we obtain

$$\mathbb{E}[\|q(X; x_t)\|^2 \mid x_t] = O(h^{2-d}).$$

To quantify the empirical deviation from this population moment, write

$$\frac{1}{n} \sum_{i=1}^n \|q_i(x_t)\|^2 = \mathbb{E}[\|q(X; x_t)\|^2 \mid x_t] + O_p\Big(n^{-1/2} h^{2-3d/2}\Big) = O_p(h^{2-d}),$$

where the last equality follows under the standard asymptotic regime ($nh^d \to \infty$).

Combining this with the finite-population identity yields

$$\begin{aligned} & T_1 \\ &:= \mathbb{E}[\|\bar{q}_t\|^2 \mid x_t, \mathcal{X}, \mathcal{G}_n^{\text{local}}] \\ &\lesssim \frac{n-m}{m(n-1)} O_p(h^{2-d}) + \Big(1 - \frac{n-m}{m(n-1)}\Big) \|k_n(x_t)\|^2 \\ &\le \frac{n-m}{m(n-1)} O_p(h^{2-d}) + \Big(1 - \frac{n-m}{m(n-1)}\Big) \Big(h^4 \alpha_j^2 \|\delta_t\|^2 + O_p\Big(h^8 + h^4 \tfrac{\log n}{nh^{d+4}}\Big) \|\delta_t\|^2 + O_p\Big(h^8 + h^4 \tfrac{\log n}{nh^{d+2}}\Big)\Big), \end{aligned}$$

where the last line follows from Lemma 9. □

Proof of Lemma 8. Define

$$N_h(x_t) := \#\{i : \|X_i - x_t\| \le h\}.$$

From the elementary inequality used in Lemma 1 we have, for any i,

$$\|q_i(x_t)\| \le \frac{\|X_i - x_t\| \exp(-\|X_i - x_t\|^2/(2h^2))}{\hat{p}_h^*(x_t)} \le \frac{he^{-1/2}}{\hat{p}_h^*(x_t)}.$$

Hence the event $\{\|q_i(x_t)\| > C\}$ implies

$$\hat{p}_h^*(x_t) < \frac{he^{-1/2}}{C_*} \quad \Longrightarrow \quad e^{-1/2} \frac{N_h(x_t)}{n} < \frac{he^{-1/2}}{C_*} \quad \Longrightarrow \quad N_h(x_t) < \frac{nh}{C_*}.$$

Thus

$$\{\|q_i(x_t)\| > C_*\} \subseteq \Big\{N_h(x_t) < \frac{nh}{C_*}\Big\}.$$

Let $\theta(x_t; h) = \Pr(\|X - x_t\| \le h) = p(x_t) V_d h^d + O(h^{d+2})$, and on the local good event $p(x_t) \ge p_{\min,j} > 0$, so for small enough h,

$$\theta(x_t; h) \ge \tfrac{1}{2} p_{\min,j} V_d h^d.$$

Therefore $\mathbb{E}[N_h(x_t)] = n\theta(x_t;h) \gtrsim nh^d$. The threshold nh/C_* scales like nh^d with a strictly smaller constant for our choice of C_* (Lemma 1 gives $C_* \asymp h^{1-d}/(p_{\min,j}V_d)$). Hence N_h falling below nh/C_* is a nontrivial multiplicative deviation below its mean. By the multiplicative Chernoff bound for a binomial variable there exist constants $c, c' > 0$ (depending on $p_{\min,j}, V_d, C_*$) such that

$$\Pr\left(N_h(x_t) < \frac{nh}{C_*}\right) \le \Pr\left(N_h(x_t) < c_0 nh^d\right) \le \exp(-c\,\mathbb{E}[N_h(x_t)]) \le \exp(-c' nh^d).$$

Define

$$\delta_n := \Pr\left(\|q(X;x_t)\| > C_*\right) \le \exp(-c' nh^d). \tag{22}$$

By Cauchy–Schwarz,

$$\|\mathbb{E}[q(X;x_t)\mathbf{1}_{\{\|q\|>C_*\}}]\| \le \sqrt{\mathbb{E}\|q(X;x_t)\|^2}\,\sqrt{\Pr(\|q\| > C_*)} \le C_1\, h^{1-d/2}\,\exp(-c' nh^d/2), \tag{23}$$

for some constant $C_1 > 0$ on $\mathcal{G}_n^{\text{local}}$.

Write $b(X;x_t) = (C_*/\|q\| - 1)q(X;x_t)\mathbf{1}_{\{\|q\|>C_*\}}$ and observe $\|b\| \le \|q\|\mathbf{1}_{\{\|q\|>C_*\}}$. Then, on $\mathcal{G}_n^{\text{local}}$

$$\mathbb{E}\|b(X;x_t)\|^2 \le \mathbb{E}(\|q(X;x_t)\|^2\mathbf{1}_{\{\|q\|>C_*\}}) \le \sqrt{\mathbb{E}\|q(X;x_t)\|^4}\,\sqrt{\Pr(\|q\| > C_*)} \le C_2\, h^{2-3d/2}\,\exp(-c' nh^d/2)$$

for some constant $C_2 > 0$, since $\mathbb{E}\|q\|^4 = O_p(h^{4-3d})$ by an argument similar to the one in (21). Note that by Assumption 2, $nh^d \ge C\log(n)$ for a sufficiently large constant $C > 0$ such that $\exp(-c' nh^d/2) \le n^{-9/d} \le (h^d)^{9/d} = h^9$ and hence

$$h^{2-3d/2}\,\exp(-c' nh^d/2) \le C' h^9.$$

□

Proof of Lemma 9. Invoke Lemma 4 and Lemma 2: on $\overline{B}(\mu_j, r_j)$ (hence on $\mathcal{G}_n^{\text{local}}$) the uniform stochastic expansions

$$\sup_{x\in\overline{B}(\mu_j,r_j)} \|\nabla\hat{\ell}_h(x) - \nabla\log p(x)\| = O_p\Big(h^2 + \sqrt{\tfrac{\log n}{nh^{d+2}}}\Big),$$

$$\sup_{x\in\overline{B}(\mu_j,r_j)} \|\nabla^2\hat{\ell}_h(x) - \nabla^2\log p(x)\| = O_p\Big(h^2 + \sqrt{\tfrac{\log n}{nh^{d+4}}}\Big)$$

hold. Using these, expand $k_n(x_t) = h^2\nabla\log\hat{p}_h(x_t)$ by Taylor's theorem about μ_j with integral-form remainder and substitute the uniform Hessian approximation together with Lemma 2 (which gives $\nabla^2\log p(\cdot) \preceq -\frac{\alpha_j}{2}I$ on $\overline{B}(\mu_j, r_j)$). One obtains, on $\mathcal{G}_n^{\text{local}}$,

$$k_n(x_t) = h^2\Big(\int_0^1 \nabla^2\log\hat{p}_h(\mu_j + s\delta_t)\,ds\Big)\delta_t + O_p\Big(h^4 + h^2\sqrt{\tfrac{\log n}{nh^{d+2}}}\Big)$$

which implies that

$$\begin{aligned}
\|k_n(x_t)\|^2 &\le 2h^4\left\|\int_0^1 \nabla^2\log\hat{p}_h(\mu_j + s\delta_t)\,ds\right\|^2\|\delta_t\|^2 + O_p\Big(h^8 + h^4\tfrac{\log n}{nh^{d+2}}\Big)\\
&\le 2h^4\left\|\int_0^1 \nabla^2\log p_h(\mu_j + s\delta_t)\,ds\right\|^2\|\delta_t\|^2 + O_p\Big(h^8 + h^4\tfrac{\log n}{nh^{d+4}}\Big)\|\delta_t\|^2 + O_p\Big(h^8 + h^4\tfrac{\log n}{nh^{d+2}}\Big)\\
&\le 2h^4\alpha_j^2\|\delta_t\|^2 + O_p\Big(h^8 + h^4\tfrac{\log n}{nh^{d+4}}\Big)\|\delta_t\|^2 + O_p\Big(h^8 + h^4\tfrac{\log n}{nh^{d+2}}\Big).
\end{aligned}$$

where the first inequality follows by Cauchy-Schwarz, the second inequality uses Lemma 4 to move from the empirical Hessian $\nabla^2\log\hat{p}_h$ to the population Hessian $\nabla^2\log p_h$, and the third term uses Assumption 3. □

Proof of Lemma 10. For each $j \in [k]$, let $A_j := \mathcal{G}_n^{\text{local},j}$ and let $B := \mathcal{E}_{\text{init}}$. By Lemma 6 and Proposition 3.4, we may assume that for all $n \geq n_0$,

$$\Pr(A_j^c) \;\leq\; \frac{1}{2k}\, n^{-2} \quad (j = 1, \ldots, k), \qquad \Pr(B^c) \;\leq\; \frac{1}{2}\, n^{-2}.$$

Then, by the union bound,

$$\Pr\Big(\bigcap_{j=1}^{k} A_j \cap B\Big) = 1 - \Pr\Big(\Big(\bigcap_{j=1}^{k} A_j \cap B\Big)^c\Big) \;\geq\; 1 - \sum_{j=1}^{k} \Pr(A_j^c) - \Pr(B^c) \;\geq\; 1 - n^{-2}.$$

By definition, $\mathcal{G}_n^{\text{global}} = \bigcap_{j=1}^{k} A_j \cap B$, so $\Pr(\mathcal{G}_n^{\text{global}}) \geq 1 - n^{-2}$ for all $n \geq n_0$. □

References

Balle, B., Barthe, G., and Gaboardi, M. (2018). Privacy amplification by subsampling: Tight analyses via couplings and divergences. *Advances in neural information processing systems*, 31.

Dwork, C., Roth, A., et al. (2014). The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407.

Genovese, C. R., Perone-Pacifico, M., Verdinelli, I., and Wasserman, L. (2014). Nonparametric ridge estimation. *The Annals of Statistics*, 42(4):1511 – 1545.

Hall, R., Rinaldo, A., and Wasserman, L. (2013). Differential privacy for functions and functional data. *The Journal of Machine Learning Research*, 14(1):703–727.