Steepest Descent Density Control for Compact 3D Gaussian Splatting

Supplementary Material

A. Implementation Details

A.1. Pseudocode

We provide a reference pseudocode for our method, SteepGS, in Algorithm 1. We highlight the main differences from the original ADC in orange. The overall procedure consists of two main components. First, the algorithm estimates the splitting matrices on the fly in a mini-batch manner. Second, at regular intervals, the accumulated splitting matrices are used to decide whether to split a Gaussian point and where to place the resulting offspring. Our algorithm is designed to be general and can be integrated with other point selection criteria, such as the gradient-based strategy used in the original ADC. Finally, we note that all `for` loops in the pseudocode are executed in parallel for efficiency.

Algorithm 1 Steepest Gaussian Splatting (SteepGS)

Input: An initial point cloud of Gaussians $\boldsymbol{\theta} = \{(\boldsymbol{\theta}^{(i)}, o^{(i)})\}_{i=1}^{|\boldsymbol{\theta}|}$; A loss function $\mathcal{L}(\boldsymbol{\theta})$ associated with a training set $\mathcal{D}(\mathcal{X})$; A stepsize $\epsilon > 0$; A splitting matrix threshold $\varepsilon_{split} \leq 0$; Total number of iterations T; Densification interval T_{split}.
for each training step $t = 1, \cdots, T$ **do**
  if $t \mod T_{split} \neq 0$ **then**
    Sample a batch of data points $\boldsymbol{x} \sim \mathcal{D}(\mathcal{X})$ and compute loss function $\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{x})$.
    for each Gaussian $i = 1, \cdots, |\boldsymbol{\theta}|$ **do**
      Update each Gaussian parameters $\boldsymbol{\theta}^{(i)}, o^{(i)}$ via standard gradient descent.
      Accumulate gradients: $\boldsymbol{G}^{(i)} \leftarrow \boldsymbol{G}^{(i)} + \nabla_{\boldsymbol{\theta}^{(i)}} \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{x})$.
      Accumulate splitting matrix: $\boldsymbol{S}^{(i)} \leftarrow \boldsymbol{S}^{(i)} + \partial_{\sigma^{(i)}} \ell(\boldsymbol{\theta}, \boldsymbol{x}) \nabla^2 \sigma(\boldsymbol{\theta}^{(i)}, \boldsymbol{x})$.
    end for
  else
    for each Gaussian $i = 1, \cdots, |\boldsymbol{\theta}|$ **do**
      Obtain average gradient and splitting matrix: $\boldsymbol{G}^{(i)} \leftarrow \boldsymbol{G}^{(i)}/T_{split}$, $\boldsymbol{S}^{(i)} \leftarrow \boldsymbol{S}^{(i)}/T_{split}$.
      Compute the smallest eigenvalue and the associated eigenvector for the splitting matrix:
      $\lambda \leftarrow \lambda_{min}(\boldsymbol{S}^{(i)})$, $\boldsymbol{\delta} \leftarrow \boldsymbol{v}_{min}(\boldsymbol{S}^{(i)})$.
      if condition on $\boldsymbol{G}^{(i)}$ and $\lambda < \varepsilon_{split}$ **then**
        Replace this Gaussian with two Gaussian off-springs:
        $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} \setminus \{(\boldsymbol{\theta}^{(i)}, o^{(i)})\} \cup \{(\boldsymbol{\theta}^{(i)} + \epsilon\boldsymbol{\delta}, o^{(i)}/2), (\boldsymbol{\theta}^{(i)} - \epsilon\boldsymbol{\delta}, o^{(i)}/2)\}$
      end if
    end for
  end if
end for
Return θ

A.2. Variants

Densification with Increment Budget. Recent densification algorithms [3, 14] have shown that fixing the number or ratio of incremental points can lead to a more compact Gaussian point cloud. This corresponds to imposing a global constraint on the total number of new points, $\sum_{i\in[n]} m_i \leq 2K$, when solving the objective in Eq. 7:

$$\min \mathcal{L}(\boldsymbol{\vartheta}, \boldsymbol{w}), \quad \text{s.t.} \left\| \boldsymbol{\vartheta}_j^{(i)} - \boldsymbol{\theta}^{(i)} \right\|_2 \leq \epsilon, \sum_{j=1}^{m_i} w_j^{(i)} = 1, \sum_{i\in[n]} m_i \leq 2K, \tag{8}$$

where K is the maximum number of increased points. According to Theorem 2, the maximal loss reduction achieved by splitting the i-th Gaussian is given by $\Delta^{(i)*} \propto \lambda_{min}(\boldsymbol{S}^{(i)}(\boldsymbol{\theta}))/2$. Therefore, the optimal point selection maximizing loss

	Tank & Temple		Deep Blending		mip-NeRF 360 Outdoor			mip-NeRF 360 Indoor			
	Train	Truck	Dr. Johnson	Playroom	Bicycle	Garden	Stump	Bonsai	Counter	Kitchen	Room
3DGS	22.091	25.394	29.209	30.172	25.253	27.417	26.705	32.298	29.006	31.628	31.540
SteepGS	21.974	25.395	29.478	30.447	24.890	27.159	26.115	31.911	28.737	31.030	31.401

Table 2. Breakdown table for per-scene PNSR of 3DGS and our SteepGS.

	Tank & Temple		Deep Blending		mip-NeRF 360 Outdoor			mip-NeRF 360 Indoor			
	Train	Truck	Dr. Johnson	Playroom	Bicycle	Garden	Stump	Bonsai	Counter	Kitchen	Room
3DGS	0.813	0.882	0.901	0.907	0.766	0.867	0.908	0.773	0.942	0.928	0.919
SteepGS	0.802	0.879	0.902	0.909	0.734	0.851	0.742	0.938	0.900	0.922	0.915

Table 3. Breakdown table for per-scene SSIM of 3DGS and our SteepGS.

descent can be done by efficiently choosing Gaussians with least-K values of $\lambda_{min}(\boldsymbol{S}^{(i)})$ once the total number of Gaussians with negative $\lambda_{min}(\boldsymbol{S}^{(i)})$ surpasses K, *i.e.* $|\{i \in [n] : \lambda_{min}(\boldsymbol{S}^{(i)}) < 0\}| > K$.

Compactest Splitting Strategy. There also exists a theoretically most compact splitting strategy. Theorem 1 suggests that the optimal displacement $\boldsymbol{\mu}$ corresponds to the standard negative gradient $\nabla\mathcal{L}(\boldsymbol{\theta})$, which yields a typical $\mathcal{O}(\epsilon)$ decrease in loss at non-stationary points. In contrast, splitting introduces a summation of splitting characteristic functions, each governed by its associated splitting matrix, resulting in a cumulative effect of order $\mathcal{O}(\epsilon^2)$. This theoretical insight leads to an important implication: *a Gaussian should be split only when its gradient is small*. Otherwise, splitting introduces redundant Gaussians that offer little improvement in loss. The compactest splitting condition can be formulated as below:

$$\|\boldsymbol{G}^{(i)}\| \leq \varepsilon_{grad} \quad \text{and} \quad \lambda_{min}(\boldsymbol{S}^{(i)}) < \varepsilon_{split}, \quad \forall i \in [n],$$

where $\varepsilon_{grad} > 0$ is a chosen hyper-parameter. While this conclusion may appear to contradict the original ADC strategy, we argue that ADC actually examines the variance of the gradient by estimating $\mathbb{E}[\|\boldsymbol{G}^{(i)}\|]$, rather than the norm of its expectation, *i.e.* $\mathbb{E}[\|\boldsymbol{G}^{(i)}\|]$. Thus, our condition does not contradict the original approach, but rather complements it by offering a more principled criterion.

A.3. Eigendecomposition

In our experiments, we only take position parameters into the consideration for steepest splitting descent. This simplifies the eigendecomposition of splitting matrices to be restricted to symmetric 3×3 matrices. We can follow the method by [29] to compute the eigenvalues. The characteristic equation of a symmetric 3×3 matrix $\boldsymbol{A}$ is:

$$\det(\alpha\boldsymbol{I} - \boldsymbol{A}) = \alpha^3 - \alpha^2 \operatorname{tr}(\boldsymbol{A}) - \alpha\frac{1}{2}\left(\operatorname{tr}(\boldsymbol{A}^2) - \operatorname{tr}^2(\boldsymbol{A})\right) - \det(\boldsymbol{A}) = 0.$$

An affine change to $\boldsymbol{A}$ will simplify the expression considerably, and lead directly to a trigonometric solution. If $\boldsymbol{A} = p\boldsymbol{B} + q\boldsymbol{I}$, then $\boldsymbol{A}$ and $\boldsymbol{B}$ have the same eigenvectors, and β is an eigenvalue of $\boldsymbol{B}$ if and only if $\alpha = p\beta + q$ is an eigenvalue of $\boldsymbol{A}$. Let $q = \frac{\operatorname{tr}(\boldsymbol{A})}{3}$ and $p = \left(\operatorname{tr}\left(\frac{(\boldsymbol{A}-q\boldsymbol{I})^2}{6}\right)\right)^{1/2}$, we derive $\det(\beta\boldsymbol{I} - \boldsymbol{B}) = \beta^3 - 3\beta - \det(\boldsymbol{B}) = 0$. Substitute $\beta = 2\cos\theta$ and some algebraic simplification using the identity $\cos 3\theta = 4\cos^3\theta - 3\cos\theta$, we can obtain $\cos 3\theta = \frac{\det(\boldsymbol{B})}{2}$. Thus, the roots of characteristic equation are given by:

$$\beta = 2\cos\left(\frac{1}{3}\arccos\left(\frac{\det(\boldsymbol{B})}{2}\right) + \frac{2k\pi}{3}\right), \quad k = 0, 1, 2.$$

When $\boldsymbol{A}$ is real and symmetric, $\det(\boldsymbol{B})$ is also real and no greater than 2 in absolute value.

B. More Experiment Results

Metrics Breakdown. Tables 2, 3, 4 and 5 provide breakdown numerical evaluations of PSNR, SSIM, LPIPS, and the number of points for both our method and the original adaptive density control. The results demonstrate that our method achieves performance comparable to the original densification across all scenes. Notably, in the `Playroom` and `Dr. Johnson` scenes, our method outperforms the original adaptive density control while utilizing only half the number of points.

	Tank & Temple		Deep Blending		mip-NeRF 360 Outdoor			mip-NeRF 360 Indoor			
	Train	Truck	Dr. Johnson	Playroom	Bicycle	Garden	Stump	Bonsai	Counter	Kitchen	Room
3DGS	0.207	0.147	0.244	0.244	0.210	0.107	0.215	0.203	0.200	0.126	0.219
SteepGS	0.230	0.160	0.251	0.250	0.268	0.142	0.271	0.211	0.217	0.137	0.233

Table 4. Breakdown table for per-scene LPIPS of 3DGS and our SteepGS.

	Tank & Temple		Deep Blending		mip-NeRF 360 Outdoor			mip-NeRF 360 Indoor			
	Train	Truck	Dr. Johnson	Playroom	Bicycle	Garden	Stump	Bonsai	Counter	Kitchen	Room
3DGS	1088197	2572172	3316036	2320830	6074705	5845401	4863462	1260017	1195896	1807771	1550152
SteepGS	530476	1387065	1485567	1107604	2900640	2195185	3175021	746163	591508	922717	710212

Table 5. Breakdown table for the number of points densified by 3DGS and our SteepGS.

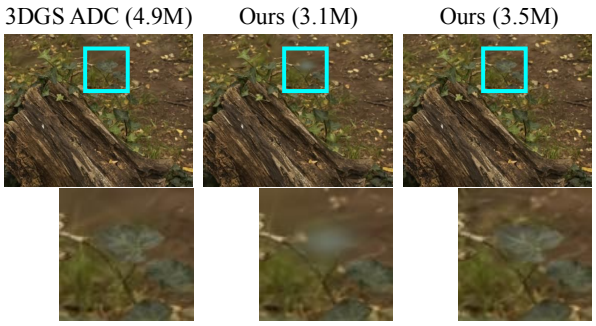

Figure 6. Improved visual quality of our method after more steps of Gaussian splitting.

	Bicycle	Garden	Stump
3DGS	25.25	27.42	26.70
Ours	24.89	27.16	26.11
Ours (more steps)	25.23	27.38	26.65

Table 6. Improved performance of our method evaluated in PSNR after more steps of Gaussian splitting.

More Visualizations. Fig. 7 visualizes the points selected for densification in four scenes. It can be observed that our method selects fewer points by concentrating on regions with blurry under-reconstructed areas. In contrast, the original adaptive density control performs more densifications on high-frequency details, which is less likely to effectively enhance rendering quality. These findings validate that our method conserves computational resources by directing densification toward areas that result in the steepest descent in rendering loss.

More Metrics and Compared Methods. In addition to the compared methods in the main text, we test two more baselines: Compact-3DGS [16] and LP-3DGS [45]. We also include elapsed time on GPU for training, mean and peak GPU memory usage for training, and rendering FPS[4] as additional metrics. Table 7 presents the comparison results evaluated on MipNeRF360, Temple&Tanks, and Deep Blending datasets. Although Compact-3DGS and LP-3DGS yield fewer points in the final results, our method achieves better metrics in PSNR and significantly reduces training time on GPUs. Moreover, our method consistently decreases GPU memory usage and improves rendering FPS compared to the original 3DGS ADC, performing on par with the two newly compared methods.

Improved Performance. Readers might feel curious if our method could achieve even more closer performance to that of the original 3DGS ADC. In our main experiments, to ensure fair comparisons, we reuse the hyper-parameters of ADC. However, we found that extending the densification iterations to 25K and the total training steps to 40K on some MipNeRF360 scenes allows our method to achieve better performance and further mitigates the blurriness observed in the rendered images. As a reference, Table 6 demonstrates performance improvements with more densification iterations. Figure 6 shows reduced blurriness in the `stump` scene.

[4]We observed that measuring FPS can be inconsistent, and the values reported in the table should be considered as a reference.

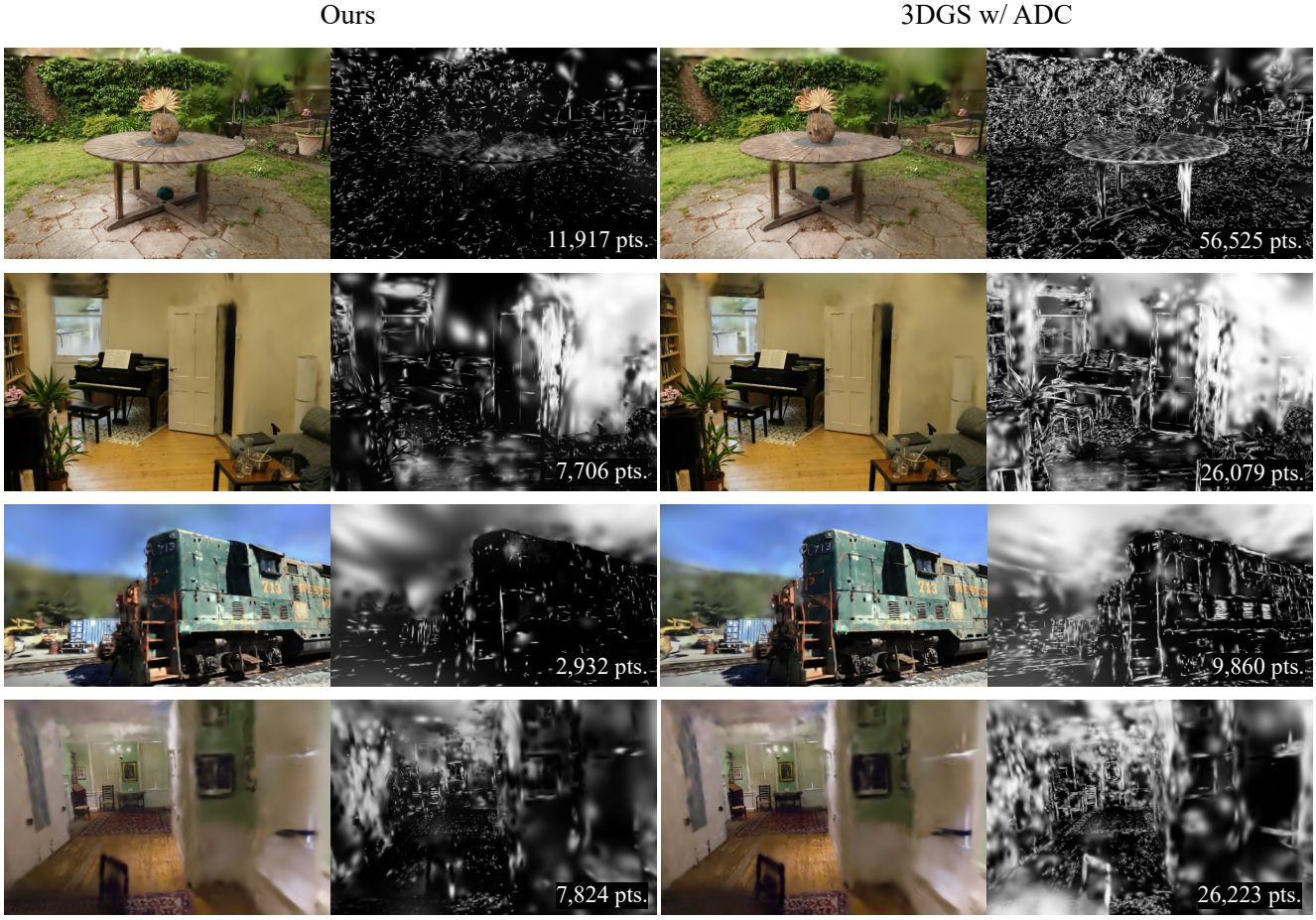

Figure 7. More visualizations of splitting points. We compare the number of points split by our proposed method and the original ADC.

C. Theory

C.1. Notations and Setup

To begin with, we re-introduce our notations and the problem setup more rigorously. We abstract each Gaussian as a function $\sigma_\Pi(\boldsymbol{x}; \boldsymbol{\theta}^{(i)}) : \Theta \times \mathcal{X} \to \mathcal{O}$ where $\boldsymbol{\theta}^{(i)} \in \Theta$ are parameters encapsulating mean, covariance, density, SH coefficients, $(\Pi, \boldsymbol{x}) \in \mathcal{X}$ denote the camera transformations and the 2D-pixel coordinates respectively, and output includes density and RGB color in space $\mathcal{O}$. Further on, we assign the input space a probability measure $\mathcal{D}(\mathcal{X})$. We combine α-blending and the photometric loss as a single function $\ell(\cdot) : \mathbb{P}(\mathcal{O}) \mapsto \mathbb{R}$, where $\mathbb{P}(\mathcal{O})$ denotes the entire output space, i.e., all multisets whose elements are in the output space $\mathcal{O}$. Suppose the scene has n Gaussians, then we denote the all parameters as $\boldsymbol{\theta} = \{\boldsymbol{\theta}^{(i)}\}_{i=1}^n$ for shorthand and the total loss function can be expressed as:

$$\mathcal{L}(\boldsymbol{\theta}) = \mathbb{E}_{\Pi, \boldsymbol{x} \sim \mathcal{D}(\mathcal{X})}[\ell(\sigma_\Pi(\boldsymbol{x}; \boldsymbol{\theta}^{(1)}), \cdots, \sigma_\Pi(\boldsymbol{x}; \boldsymbol{\theta}^{(n)}))]. \tag{9}$$

Now our goal is to split each Gaussian into m_i off-springs. We denote the parameters of the i-th Gaussian's off-springs as $\boldsymbol{\vartheta}^{(i)} = \{\boldsymbol{\vartheta}_j^{(i)}\}_{j=1}^{m_i}$, where $\boldsymbol{\vartheta}_j^{(i)}$ is the j-th off-spring of the i-th Gaussian and assign it a group of reweighting coefficients $\boldsymbol{w}^{(i)} = \{w_j^{(i)}\}_{j=1}^{m_i}$ to over-parameterize the original Gaussian as: $\sum_{j=1}^{m_i} w_j^{(i)} \sigma_\Pi(\boldsymbol{x}; \boldsymbol{\vartheta}_j^{(i)})$ such that $\sum_{j=1}^{m_i} w_j^{(i)} = 1$ for every $i \in [n]$ We collect parameters of all the new Gaussians as $\boldsymbol{\vartheta} = \{\boldsymbol{\vartheta}^{(i)}\}_{i=1}^n$, and reweighting coefficients as $\boldsymbol{w} = \{\boldsymbol{w}^{(i)}\}_{i=1}^n$, for

MipNeRF360								
	# Points ↓	PSNR ↑	SSIM ↑	LPIPS ↓	GPU elapse ↓	mean GPU mem. ↓	peak GPU mem. ↓	FPS ↑
3DGS	3.339 M	29.037	0.872	0.183	1550.925 s	10.262 GB	12.110 GB	179
LP-3DGS	1.303 M	28.640	0.865	0.198	1177.648 s	10.027 GB	12.458 GB	350
Compact-3DGS	1.310 M	28.504	0.856	0.208	4063.203 s	7.274 GB	9.044 GB	98
SteepGS (Ours)	1.606 M	28.734	0.857	0.211	1051.276 s	7.597 GB	8.957 GB	252
Tank & Temple								
	# Points ↓	PSNR ↑	SSIM ↑	LPIPS ↓	GPU elapse ↓	mean GPU mem. ↓	peak GPU mem. ↓	FPS ↑
3DGS	1.830 M	23.743	0.848	0.177	803.542 s	5.193 GB	6.241 GB	248
LP-3DGS	0.671 M	23.424	0.839	0.197	1021.806 s	5.045 GB	6.489 GB	150
Compact-3DGS	0.836 M	23.319	0.835	0.200	1255.748 s	3.802 GB	4.774 GB	357
SteepGS (Ours)	0.958 M	23.684	0.840	0.194	539.048 s	4.701 GB	5.607 GB	343
Deep Blending								
	# Points ↓	PSNR ↑	SSIM ↑	LPIPS ↓	GPU elapse ↓	mean GPU mem. ↓	peak GPU mem. ↓	FPS ↑
3DGS	2.818 M	29.690	0.904	0.244	1429.878 s	8.668 GB	10.218 GB	187
LP-3DGS	0.861 M	29.764	0.906	0.249	1697.793 s	8.354 GB	10.115 GB	134
Compact-3DGS	1.054 M	29.896	0.905	0.255	1861.897 s	6.332 GB	8.026 GB	312
SteepGS (Ours)	1.296 M	29.963	0.905	0.250	956.536 s	5.928 GB	9.506 GB	280

Table 7. Comparison with LP-3DGS [45] and Compact-3DGS [16] baselines on MipNeRF360, Tank & Temple, and Deep Blending datasets. Additional metrics: GPU elapsed time for training, mean & peak GPU memory usage, and FPS are included.

shorthand. With newly added Gaussians, the augmented loss function becomes:

$$\mathcal{L}(\boldsymbol{\vartheta}, \boldsymbol{w}) = \mathbb{E}_{\Pi, \boldsymbol{x}\sim\mathcal{D}(\mathcal{X})}\left[\ell\left(\sum_{j=1}^{m_1} w_j^{(1)}\sigma_\Pi(\boldsymbol{x};\boldsymbol{\vartheta}_j^{(1)}), \cdots, \sum_{j=1}^{m_n} w_j^{(n)}\sigma_\Pi(\boldsymbol{x};\boldsymbol{\vartheta}_j^{(n)})\right)\right]. \tag{10}$$

C.2. Main Results

Proof of Theorem 1. We define $\boldsymbol{\mu}^{(i)}$ as the average displacement on $\boldsymbol{\theta}^{(i)}$: $\boldsymbol{\mu}^{(i)} = (\sum_{j=1}^{m_i} w_j^{(i)}\boldsymbol{\vartheta}_j^{(i)} - \boldsymbol{\theta}^{(i)})/\epsilon$ and $\boldsymbol{\delta}_j^{(i)} = (\boldsymbol{\vartheta}_j^{(i)} - \boldsymbol{\theta}^{(i)})/\epsilon - \boldsymbol{\mu}^{(i)}$ as offset additional to $\boldsymbol{\mu}^{(i)}$ for the j-th off-spring. It is obvious that:

$$\begin{aligned}\sum_{j=1}^{m_i} w_j^{(i)}\boldsymbol{\delta}_j^{(i)} &= \sum_{j=1}^{m_i} w_j^{(i)}\left(\frac{\boldsymbol{\vartheta}_j^{(i)} - \boldsymbol{\theta}^{(i)}}{\epsilon} - \boldsymbol{\mu}^{(i)}\right) = \frac{1}{\epsilon}\sum_{j=1}^{m_i} w_j^{(i)}\boldsymbol{\vartheta}_j^{(i)} - \frac{1}{\epsilon}\sum_{j=1}^{m_i} w_j^{(i)}\boldsymbol{\theta}^{(i)} - \sum_{j=1}^{m_i} w_j^{(i)}\boldsymbol{\mu}^{(i)} \\ &= \frac{1}{\epsilon}\sum_{j=1}^{m_i} w_j^{(i)}\boldsymbol{\vartheta}_j^{(i)} - \frac{1}{\epsilon}\boldsymbol{\theta}^{(i)} - \boldsymbol{\mu}^{(i)} = \mathbf{0}.\end{aligned} \tag{11}$$

In addition, we let $\boldsymbol{\Delta}_j^{(i)} = \boldsymbol{\mu}^{(i)} + \boldsymbol{\delta}_j^{(i)}$, and $\boldsymbol{\vartheta}_j^{(i)}$ can be written as: $\boldsymbol{\vartheta}_j^{(i)} = \boldsymbol{\theta}^{(i)} + \epsilon\boldsymbol{\Delta}_j^{(i)} = \boldsymbol{\theta}^{(i)} + \epsilon(\boldsymbol{\mu}^{(i)} + \boldsymbol{\delta}_j^{(i)})$. We define an auxiliary function: $\mathcal{L}(\boldsymbol{\theta}^{(\setminus i)}, \boldsymbol{\vartheta}^{(i)}, \boldsymbol{w}^{(i)})$ as:

$$\mathcal{L}(\boldsymbol{\theta}^{(\setminus i)}, \boldsymbol{\vartheta}^{(i)}, \boldsymbol{w}^{(i)}) = \mathbb{E}_{\Pi, \boldsymbol{x}\sim\mathcal{D}(\mathcal{X})}\left[\ell\left(\sigma_\Pi(\boldsymbol{x};\boldsymbol{\theta}^{(1)}), \cdots, \sum_{j=1}^{m_i} w_j^{(i)}\sigma_\Pi(\boldsymbol{x};\boldsymbol{\vartheta}_j^{(i)}), \cdots, \sigma_\Pi(\boldsymbol{x};\boldsymbol{\theta}^{(n)})\right)\right], \tag{12}$$

which only splits the i-th Gaussian $\boldsymbol{\theta}^{(i)}$ as $\boldsymbol{\vartheta}^{(i)}$. By Lemma 6, we have that:

$$(\mathcal{L}(\boldsymbol{\vartheta}, \boldsymbol{w}) - \mathcal{L}(\boldsymbol{\theta})) = \sum_{i=1}^{n} \left(\mathcal{L}(\boldsymbol{\theta}^{(\setminus i)}, \boldsymbol{\vartheta}^{(i)}, \boldsymbol{w}^{(i)}) - \mathcal{L}(\boldsymbol{\theta}) \right) + \frac{\epsilon^2}{2} \sum_{\substack{i,i' \in [n] \\ i \neq i'}} \boldsymbol{\mu}^{(i)\top} \partial^2_{\boldsymbol{\theta}^{(i)} \boldsymbol{\theta}^{(i')}} \mathcal{L}(\boldsymbol{\theta}) \boldsymbol{\mu}^{(i')} + \mathcal{O}(\epsilon^3). \tag{13}$$

By Lemma 7, we have:

$$\mathcal{L}(\boldsymbol{\theta}^{(\setminus i)}, \boldsymbol{\vartheta}^{(i)}, \boldsymbol{w}^{(i)}) - \mathcal{L}(\boldsymbol{\theta}) = \epsilon \partial_{\boldsymbol{\theta}^{(i)}} \mathcal{L}(\boldsymbol{\theta})^\top \boldsymbol{\mu}^{(i)} + \frac{\epsilon^2}{2} \boldsymbol{\mu}^{(i)\top} \partial^2_{\boldsymbol{\theta}^{(i)} \boldsymbol{\theta}^{(i)}} \mathcal{L}(\boldsymbol{\theta}) \boldsymbol{\mu}^{(i)} \tag{14}$$

$$+ \frac{\epsilon^2}{2} \sum_{j=1}^{m_i} w_j^{(i)} \boldsymbol{\delta}_j^{(i)\top} \boldsymbol{S}^{(i)}(\boldsymbol{\theta}) \boldsymbol{\delta}_j^{(i)} + \mathcal{O}(\epsilon^3). \tag{15}$$

Let $\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}^{(1)} & \cdots & \boldsymbol{\mu}^{(n)} \end{bmatrix}$ concatenate the average displacement on all Gaussians. Combining Eq. 13 and Eq. 14, we can conclude:

$$\begin{aligned}
(\mathcal{L}(\boldsymbol{\vartheta}, \boldsymbol{w}) - \mathcal{L}(\boldsymbol{\theta})) &= \sum_{i=1}^{n} \left[\epsilon \partial_{\boldsymbol{\theta}^{(i)}} \mathcal{L}(\boldsymbol{\theta})^\top \boldsymbol{\mu}^{(i)} + \frac{\epsilon^2}{2} \boldsymbol{\mu}^{(i)\top} \partial^2_{\boldsymbol{\theta}^{(i)} \boldsymbol{\theta}^{(i)}} \mathcal{L}(\boldsymbol{\theta}) \boldsymbol{\mu}^{(i)} + \frac{\epsilon^2}{2} \sum_{j=1}^{m_i} w_j^{(i)} \boldsymbol{\delta}_j^{(i)\top} \boldsymbol{S}^{(i)}(\boldsymbol{\theta}) \boldsymbol{\delta}_j^{(i)} \right] \\
&\quad + \frac{\epsilon^2}{2} \sum_{\substack{i,i' \in [n] \\ i \neq i'}} \boldsymbol{\mu}^{(i)\top} \partial^2_{\boldsymbol{\theta}^{(i)} \boldsymbol{\theta}^{(i')}} \mathcal{L}(\boldsymbol{\theta}) \boldsymbol{\mu}^{(i')} + \mathcal{O}(\epsilon^3) \\
&= \epsilon \sum_{i=1}^{n} \partial_{\boldsymbol{\theta}^{(i)}} \mathcal{L}(\boldsymbol{\theta})^\top \boldsymbol{\mu}^{(i)} + \frac{\epsilon^2}{2} \sum_{i,i' \in [n]} \boldsymbol{\mu}^{(i)\top} \partial^2_{\boldsymbol{\theta}^{(i)} \boldsymbol{\theta}^{(i')}} \mathcal{L}(\boldsymbol{\theta}) \boldsymbol{\mu}^{(i')} \\
&\quad + \frac{\epsilon^2}{2} \sum_{j=1}^{m_i} w_j^{(i)} \boldsymbol{\delta}_j^{(i)\top} \boldsymbol{S}^{(i)}(\boldsymbol{\theta}) \boldsymbol{\delta}_j^{(i)} + \mathcal{O}(\epsilon^3) \\
&= \epsilon \nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta})^\top \boldsymbol{\mu} + \frac{\epsilon^2}{2} \boldsymbol{\mu}^\top \nabla^2 \mathcal{L}(\boldsymbol{\theta}) \boldsymbol{\mu} + \frac{\epsilon^2}{2} \sum_{j=1}^{m_i} w_j^{(i)} \boldsymbol{\delta}_j^{(i)\top} \boldsymbol{S}^{(i)}(\boldsymbol{\theta}) \boldsymbol{\delta}_j^{(i)} + \mathcal{O}(\epsilon^3),
\end{aligned}$$

as desired. □

Proof of Theorem 2. By standard variational characterization, we have the following lower bound:

$$\Delta^{(i)}(\boldsymbol{\delta}^{(i)}, \boldsymbol{w}^{(i)}; \boldsymbol{\theta}) := \frac{\epsilon^2}{2} \sum_{j=1}^{m_i} w_j^{(i)} \boldsymbol{\delta}_j^{(i)\top} \boldsymbol{S}^{(i)}(\boldsymbol{\theta}) \boldsymbol{\delta}_j^{(i)} \geq \frac{\epsilon^2}{2} \sum_{j=1}^{m_i} w_j^{(i)} \lambda_{min}(\boldsymbol{S}^{(i)}(\boldsymbol{\theta})) = \frac{\epsilon^2}{2} \lambda_{min}(\boldsymbol{S}^{(i)}(\boldsymbol{\theta})),$$

subject to $\|\boldsymbol{\delta}_j^{(i)}\| \leq 1$. The equality holds only if $\boldsymbol{\delta}_j^{(i)}$ equals to the smallest eigenvector of $\boldsymbol{S}^{(i)}(\boldsymbol{\theta})$.

Hence, there is no decrease on the loss if $\lambda_{min}(\boldsymbol{S}^{(i)}(\boldsymbol{\theta})) \geq 0$. Otherwise, we can simply choose $m_i = 2$, $w_1^{(i)} = w_2^{(i)} = 1/2$, $\boldsymbol{\delta}_1^{(i)} = \boldsymbol{v}_{min}(\boldsymbol{S}^{(i)}(\boldsymbol{\theta}))$, and $\boldsymbol{\delta}_2^{(i)} = -\boldsymbol{v}_{min}(\boldsymbol{S}^{(i)}(\boldsymbol{\theta}))$ to achieve this lower bound. □

C.3. Auxiliary Results

Lemma 3. *The following equalities hold for $\mathcal{L}(\boldsymbol{\theta})$ for every $i \in [n]$*

$$\begin{aligned}
\partial_{\boldsymbol{\theta}^{(i)}} \mathcal{L}(\boldsymbol{\theta}) &= \mathbb{E}_{\Pi, \boldsymbol{x} \sim \mathcal{D}(\mathcal{X})} \left[\partial_{\sigma^{(i)}} \ell \left(\sigma_\Pi(\boldsymbol{x}; \boldsymbol{\theta}^{(1)}), \cdots, \sigma_\Pi(\boldsymbol{x}; \boldsymbol{\theta}^{(n)}) \right) \nabla \sigma_\Pi(\boldsymbol{x}; \boldsymbol{\theta}^{(i)}) \right], \\
\partial^2_{\boldsymbol{\theta}^{(i)} \boldsymbol{\theta}^{(i)}} \mathcal{L}(\boldsymbol{\theta}) &= \boldsymbol{T}^{(i)}(\boldsymbol{\theta}) + \boldsymbol{S}^{(i)}(\boldsymbol{\theta}), \\
\partial^2_{\boldsymbol{\theta}^{(i)} \boldsymbol{\theta}^{(i')}} \mathcal{L}(\boldsymbol{\theta}) &= \mathbb{E}_{\Pi, \boldsymbol{x} \sim \mathcal{D}(\mathcal{X})} \left[\partial^2_{\sigma^{(i)} \sigma^{(i')}} \ell \left(\sigma_\Pi(\boldsymbol{x}; \boldsymbol{\theta}^{(1)}), \cdots, \sigma_\Pi(\boldsymbol{x}; \boldsymbol{\theta}^{(n)}) \right) \nabla \sigma_\Pi(\boldsymbol{x}; \boldsymbol{\theta}^{(i)}) \nabla \sigma_\Pi(\boldsymbol{x}; \boldsymbol{\theta}^{(i')})^\top \right], \forall i' \in [n], i' \neq i,
\end{aligned}$$

where $\boldsymbol{T}^{(i)}(\boldsymbol{\theta}) = \mathbb{E}_{\Pi, \boldsymbol{x} \sim \mathcal{D}(\mathcal{X})} \left[\partial^2_{\sigma^{(i)} \sigma^{(i)}} \ell \left(\sigma_\Pi(\boldsymbol{x}; \boldsymbol{\theta}^{(1)}), \cdots, \sigma_\Pi(\boldsymbol{x}; \boldsymbol{\theta}^{(n)}) \right) \nabla \sigma_\Pi(\boldsymbol{x}; \boldsymbol{\theta}^{(i)}) \nabla \sigma_\Pi(\boldsymbol{x}; \boldsymbol{\theta}^{(i)})^\top \right]$.

Proof. The gradient of $\mathcal{L}(\boldsymbol{\theta})$ is proved via simple chain rule. And then

$$\begin{aligned}
&\partial^2_{\boldsymbol{\theta}^{(i)}\boldsymbol{\theta}^{(i')}}\mathcal{L}(\boldsymbol{\theta}) = \partial_{\boldsymbol{\theta}^{(i')}}\left[\partial_{\boldsymbol{\theta}^{(i)}}\mathcal{L}(\boldsymbol{\theta})\right] \\
&= \mathbb{E}_{\Pi,\boldsymbol{x}\sim\mathcal{D}(\mathcal{X})}\left[\partial^2_{\sigma^{(i)}\sigma^{(i')}}\ell\left(\sigma_\Pi(\boldsymbol{x};\boldsymbol{\theta}^{(1)}),\cdots,\sigma_\Pi(\boldsymbol{x};\boldsymbol{\theta}^{(n)})\right)\nabla\sigma_\Pi(\boldsymbol{x};\boldsymbol{\theta}^{(i)})\nabla\sigma_\Pi(\boldsymbol{x};\boldsymbol{\theta}^{(i')})^\top\right] \\
&\quad + \mathbb{E}_{\Pi,\boldsymbol{x}\sim\mathcal{D}(\mathcal{X})}\left[\partial_{\sigma^{(i)}}\ell\left(\sigma_\Pi(\boldsymbol{x};\boldsymbol{\theta}^{(1)}),\cdots,\sigma_\Pi(\boldsymbol{x};\boldsymbol{\theta}^{(n)})\right)\partial_{\boldsymbol{\theta}^{(i')}}\nabla\sigma_\Pi(\boldsymbol{x};\boldsymbol{\theta}^{(i)})\right].
\end{aligned}$$

When $i = i'$, $\partial_{\boldsymbol{\theta}^{(i')}}\nabla\sigma_\Pi(\boldsymbol{x};\boldsymbol{\theta}^{(i)}) = \nabla^2\sigma_\Pi(\boldsymbol{x};\boldsymbol{\theta}^{(i)})$, henceforth:

$$\partial^2_{\boldsymbol{\theta}^{(i)}\boldsymbol{\theta}^{(i)}}\mathcal{L}(\boldsymbol{\theta}) = \boldsymbol{T}^{(i)}(\boldsymbol{\theta}) + \boldsymbol{S}^{(i)}(\boldsymbol{\theta}).$$

Otherwise, $\partial_{\boldsymbol{\theta}^{(i')}}\nabla\sigma_\Pi(\boldsymbol{x};\boldsymbol{\theta}^{(i)}) = \mathbf{0}$, and thus:

$$\partial^2_{\boldsymbol{\theta}^{(i)}\boldsymbol{\theta}^{(i')}}\mathcal{L}(\boldsymbol{\theta}) = \mathbb{E}_{\Pi,\boldsymbol{x}\sim\mathcal{D}(\mathcal{X})}\left[\partial^2_{\sigma^{(i)}\sigma^{(i')}}\ell\left(\sigma_\Pi(\boldsymbol{x};\boldsymbol{\theta}^{(1)}),\cdots,\sigma_\Pi(\boldsymbol{x};\boldsymbol{\theta}^{(n)})\right)\nabla\sigma_\Pi(\boldsymbol{x};\boldsymbol{\theta}^{(i)})\nabla\sigma_\Pi(\boldsymbol{x};\boldsymbol{\theta}^{(i')})^\top\right],$$

all as desired. □

Lemma 4. *The following equalities hold for $\mathcal{L}(\boldsymbol{\vartheta},\boldsymbol{w})$ at $\epsilon = 0$:*

$$\partial_{\boldsymbol{\vartheta}_j^{(i)}}\mathcal{L}(\boldsymbol{\vartheta},\boldsymbol{w})\Big|_{\epsilon=0} = w_j^{(i)}\partial_{\boldsymbol{\theta}^{(i)}}\mathcal{L}(\boldsymbol{\theta}), \quad \forall i\in[n], j\in[m_i], \tag{16}$$

$$\partial^2_{\boldsymbol{\vartheta}_j^{(i)}\boldsymbol{\vartheta}_j^{(i)}}\mathcal{L}(\boldsymbol{\vartheta},\boldsymbol{w})\Big|_{\epsilon=0} = w_j^{(i)}\boldsymbol{S}^{(i)}(\boldsymbol{\theta}) + w_j^{(i)2}\boldsymbol{T}^{(i)}(\boldsymbol{\theta}), \quad \forall i\in[n], j\in[m_i], \tag{17}$$

$$\partial^2_{\boldsymbol{\vartheta}_j^{(i)}\boldsymbol{\vartheta}_{j'}^{(i)}}\mathcal{L}(\boldsymbol{\vartheta},\boldsymbol{w})\Big|_{\epsilon=0} = w_j^{(i)}w_{j'}^{(i)}\boldsymbol{T}^{(i)}(\boldsymbol{\theta}), \quad \forall i\in[n], j,j'\in[m_i], j\neq j', \tag{18}$$

$$\partial^2_{\boldsymbol{\vartheta}_j^{(i)},\boldsymbol{\vartheta}_{j'}^{(i')}}\mathcal{L}(\boldsymbol{\vartheta},\boldsymbol{w})\Big|_{\epsilon=0} = w_j^{(i)}w_{j'}^{(i')}\partial^2_{\boldsymbol{\theta}^{(i)}\boldsymbol{\theta}^{(i')}}\mathcal{L}(\boldsymbol{\theta}), \quad \forall i,i'\in[n], i\neq i', j\in[m_i], j'\in[m_{i'}], \tag{19}$$

where $\boldsymbol{T}^{(i)}(\boldsymbol{\theta}) = \mathbb{E}_{\Pi,\boldsymbol{x}\sim\mathcal{D}(\mathcal{X})}\left[\partial^2_{\sigma^{(i)}\sigma^{(i)}}\ell\left(\sigma_\Pi(\boldsymbol{x};\boldsymbol{\theta}^{(1)}),\cdots,\sigma_\Pi(\boldsymbol{x};\boldsymbol{\theta}^{(n)})\right)\nabla\sigma_\Pi(\boldsymbol{x};\boldsymbol{\theta}^{(i)})\nabla\sigma_\Pi(\boldsymbol{x};\boldsymbol{\theta}^{(i)})^\top\right]$ is as defined in Lemma 3.

Proof. Let $\widetilde{\sigma}_\Pi(\boldsymbol{x};\boldsymbol{\vartheta}^{(i)}) = \sum_{j=1}^{m_i} w_j^{(i)}\sigma_\Pi(\boldsymbol{x};\boldsymbol{\vartheta}_j^{(i)})$ and we can express $\mathcal{L}(\boldsymbol{\vartheta},\boldsymbol{w})$ as:

$$\mathcal{L}(\boldsymbol{\vartheta},\boldsymbol{w}) = \mathbb{E}_{\Pi,\boldsymbol{x}\sim\mathcal{D}(\mathcal{X})}\left[\ell\left(\widetilde{\sigma}_\Pi(\boldsymbol{x};\boldsymbol{\vartheta}^{(1)}),\cdots,\widetilde{\sigma}_\Pi(\boldsymbol{x};\boldsymbol{\vartheta}^{(n)})\right)\right]. \tag{20}$$

To take derivatives of $\mathcal{L}(\boldsymbol{\vartheta},\boldsymbol{w})$, we leverage the chain rule. For every $i\in[n], j\in[m_i]$:

$$\begin{aligned}
\partial_{\boldsymbol{\vartheta}_j^{(i)}}\mathcal{L}(\boldsymbol{\vartheta},\boldsymbol{w}) &= \partial_{\boldsymbol{\vartheta}_j^{(i)}}\mathbb{E}_{\Pi,\boldsymbol{x}\sim\mathcal{D}(\mathcal{X})}\left[\ell\left(\widetilde{\sigma}_\Pi(\boldsymbol{x};\boldsymbol{\vartheta}^{(1)}),\cdots,\widetilde{\sigma}_\Pi(\boldsymbol{x};\boldsymbol{\vartheta}^{(n)})\right)\right] \\
&= \mathbb{E}_{\Pi,\boldsymbol{x}\sim\mathcal{D}(\mathcal{X})}\left[\partial_{\boldsymbol{\vartheta}_j^{(i)}}\ell\left(\widetilde{\sigma}_\Pi(\boldsymbol{x};\boldsymbol{\vartheta}^{(1)}),\cdots,\widetilde{\sigma}_\Pi(\boldsymbol{x};\boldsymbol{\vartheta}^{(n)})\right)\right] \\
&= \mathbb{E}_{\Pi,\boldsymbol{x}\sim\mathcal{D}(\mathcal{X})}\left[\partial_{\sigma^{(i)}}\ell\left(\widetilde{\sigma}_\Pi(\boldsymbol{x};\boldsymbol{\vartheta}^{(1)}),\cdots,\widetilde{\sigma}_\Pi(\boldsymbol{x};\boldsymbol{\vartheta}^{(n)})\right)\partial_{\boldsymbol{\vartheta}_j^{(i)}}\widetilde{\sigma}_\Pi(\boldsymbol{x};\boldsymbol{\vartheta}^{(i)})\right] \\
&= w_j^{(i)}\,\mathbb{E}_{\Pi,\boldsymbol{x}\sim\mathcal{D}(\mathcal{X})}\left[\partial_{\sigma^{(i)}}\ell\left(\widetilde{\sigma}_\Pi(\boldsymbol{x};\boldsymbol{\vartheta}^{(1)}),\cdots,\widetilde{\sigma}_\Pi(\boldsymbol{x};\boldsymbol{\vartheta}^{(n)})\right)\nabla\sigma_\Pi(\boldsymbol{x};\boldsymbol{\vartheta}_j^{(i)})\right].
\end{aligned} \tag{21}$$

Since $\epsilon = 0$, we have $\boldsymbol{\vartheta}_j^{(i)} = \boldsymbol{\theta}^{(i)}$ and $\widetilde{\sigma}_\Pi(\boldsymbol{x};\boldsymbol{\vartheta}^{(i)}) = \sum_{j=1}^{m_i} w_j^{(i)}\sigma_\Pi(\boldsymbol{x};\boldsymbol{\theta}^{(i)}) = \sigma_\Pi(\boldsymbol{x};\boldsymbol{\theta})$. Hence, we can further simplify Eq. 21 as:

$$\begin{aligned}
\partial_{\boldsymbol{\vartheta}_j^{(i)}}\mathcal{L}(\boldsymbol{\vartheta},\boldsymbol{w})\Big|_{\epsilon=0} &= w_j^{(i)}\,\mathbb{E}_{\Pi,\boldsymbol{x}\sim\mathcal{D}(\mathcal{X})}\left[\partial_{\sigma^{(i)}}\ell\left(\widetilde{\sigma}_\Pi(\boldsymbol{x};\boldsymbol{\vartheta}^{(1)}),\cdots,\widetilde{\sigma}_\Pi(\boldsymbol{x};\boldsymbol{\vartheta}^{(n)})\right)\nabla\sigma_\Pi(\boldsymbol{x};\boldsymbol{\vartheta}_j^{(i)})\right]\Big|_{\epsilon=0} \\
&= w_j^{(i)}\,\mathbb{E}_{\Pi,\boldsymbol{x}\sim\mathcal{D}(\mathcal{X})}\left[\partial_{\sigma^{(i)}}\ell\left(\sigma_\Pi(\boldsymbol{x};\boldsymbol{\theta}^{(1)}),\cdots,\sigma_\Pi(\boldsymbol{x};\boldsymbol{\theta}^{(n)})\right)\nabla\sigma_\Pi(\boldsymbol{x};\boldsymbol{\theta}^{(i)})\right] \\
&= w_j^{(i)}\partial_{\boldsymbol{\theta}^{(i)}}\mathcal{L}(\boldsymbol{\theta}),
\end{aligned}$$

where the last step is due to Lemma 3.

Next we derive second-order derivatives. Taking derivatives of Eq. 21 in terms of $\boldsymbol{\vartheta}_{j'}^{(i')}$ for some $i' \in [n], j' \in [m_{i'}]$, and by chain rule:

$$\begin{aligned}
&\partial^2_{\boldsymbol{\vartheta}_j^{(i)}\boldsymbol{\vartheta}_{j'}^{(i')}}\mathcal{L}(\boldsymbol{\vartheta},\boldsymbol{w}) = \partial_{\boldsymbol{\vartheta}_{j'}^{(i')}} w_j^{(i)}\,\mathbb{E}_{\Pi,\boldsymbol{x}\sim\mathcal{D}(\mathcal{X})}\left[\partial_{\sigma^{(i)}}\ell\left(\widetilde{\sigma}_\Pi(\boldsymbol{x};\boldsymbol{\vartheta}^{(1)}),\cdots,\widetilde{\sigma}_\Pi(\boldsymbol{x};\boldsymbol{\vartheta}^{(n)})\right)\nabla\sigma_\Pi(\boldsymbol{x};\boldsymbol{\vartheta}_j^{(i)})\right]\\
&= w_j^{(i)}\,\mathbb{E}_{\Pi,\boldsymbol{x}\sim\mathcal{D}(\mathcal{X})}\left[\partial^2_{\sigma^{(i)}\sigma^{(i')}}\ell\left(\widetilde{\sigma}_\Pi(\boldsymbol{x};\boldsymbol{\vartheta}^{(1)}),\cdots,\widetilde{\sigma}_\Pi(\boldsymbol{x};\boldsymbol{\vartheta}^{(n)})\right)\nabla\sigma_\Pi(\boldsymbol{x};\boldsymbol{\vartheta}_j^{(i)})\partial_{\boldsymbol{\vartheta}_{j'}^{(i')}}\widetilde{\sigma}_\Pi(\boldsymbol{x};\boldsymbol{\vartheta}^{(i')})^\top\right]\\
&\quad + w_j^{(i)}\,\mathbb{E}_{\Pi,\boldsymbol{x}\sim\mathcal{D}(\mathcal{X})}\left[\partial_{\sigma^{(i)}}\ell\left(\widetilde{\sigma}_\Pi(\boldsymbol{x};\boldsymbol{\vartheta}^{(1)}),\cdots,\widetilde{\sigma}_\Pi(\boldsymbol{x};\boldsymbol{\vartheta}^{(n)})\right)\partial_{\boldsymbol{\vartheta}_{j'}^{(i')}}\nabla\sigma_\Pi(\boldsymbol{x};\boldsymbol{\vartheta}_j^{(i)})\right]\\
&= w_j^{(i)}w_{j'}^{(i')}\,\mathbb{E}_{\Pi,\boldsymbol{x}\sim\mathcal{D}(\mathcal{X})}\left[\partial^2_{\sigma^{(i)}\sigma^{(i')}}\ell\left(\widetilde{\sigma}_\Pi(\boldsymbol{x};\boldsymbol{\vartheta}^{(1)}),\cdots,\widetilde{\sigma}_\Pi(\boldsymbol{x};\boldsymbol{\vartheta}^{(n)})\right)\nabla\sigma_\Pi(\boldsymbol{x};\boldsymbol{\vartheta}_j^{(i)})\nabla\sigma_\Pi(\boldsymbol{x};\boldsymbol{\vartheta}_{j'}^{(i')})^\top\right]\\
&\quad + w_j^{(i)}\,\mathbb{E}_{\Pi,\boldsymbol{x}\sim\mathcal{D}(\mathcal{X})}\left[\partial_{\sigma^{(i)}}\ell\left(\widetilde{\sigma}_\Pi(\boldsymbol{x};\boldsymbol{\vartheta}^{(1)}),\cdots,\widetilde{\sigma}_\Pi(\boldsymbol{x};\boldsymbol{\vartheta}^{(n)})\right)\partial_{\boldsymbol{\vartheta}_{j'}^{(i')}}\nabla\sigma_\Pi(\boldsymbol{x};\boldsymbol{\vartheta}_j^{(i)})\right].
\end{aligned}$$

Now we discuss three scenarios:

1. When $i = i'$ and $j = j'$, $\partial_{\boldsymbol{\vartheta}_{j'}^{(i')}}\nabla\sigma_\Pi(\boldsymbol{x};\boldsymbol{\vartheta}_j^{(i)}) = \nabla^2\sigma_\Pi(\boldsymbol{x};\boldsymbol{\vartheta}_j^{(i)})$, and then

$$\begin{aligned}
\partial^2_{\boldsymbol{\vartheta}_j^{(i)}\boldsymbol{\vartheta}_j^{(i)}}\mathcal{L}(\boldsymbol{\vartheta},\boldsymbol{w}) = w_j^{(i)2}\,\mathbb{E}_{\Pi,\boldsymbol{x}\sim\mathcal{D}(\mathcal{X})}&\left[\partial^2_{\sigma^{(i)}\sigma^{(i)}}\ell\left(\widetilde{\sigma}_\Pi(\boldsymbol{x};\boldsymbol{\vartheta}^{(1)}),\cdots,\widetilde{\sigma}_\Pi(\boldsymbol{x};\boldsymbol{\vartheta}^{(n)})\right)\nabla\sigma_\Pi(\boldsymbol{x};\boldsymbol{\vartheta}_j^{(i)})\nabla\sigma_\Pi(\boldsymbol{x};\boldsymbol{\vartheta}_j^{(i)})^\top\right]\\
&+ w_j^{(i)}\,\mathbb{E}_{\Pi,\boldsymbol{x}\sim\mathcal{D}(\mathcal{X})}\left[\partial_{\sigma^{(i)}}\ell\left(\widetilde{\sigma}_\Pi(\boldsymbol{x};\boldsymbol{\vartheta}^{(1)}),\cdots,\widetilde{\sigma}_\Pi(\boldsymbol{x};\boldsymbol{\vartheta}^{(n)})\right)\nabla^2\sigma_\Pi(\boldsymbol{x};\boldsymbol{\vartheta}_j^{(i)})\right]
\end{aligned}\tag{22}$$

2. When $i = i'$ and $j \neq j'$, $\partial_{\boldsymbol{\vartheta}_{j'}^{(i')}}\nabla\sigma_\Pi(\boldsymbol{x};\boldsymbol{\vartheta}_j^{(i)}) = \mathbf{0}$, and thus

$$\partial^2_{\boldsymbol{\vartheta}_j^{(i)}\boldsymbol{\vartheta}_{j'}^{(i)}}\mathcal{L}(\boldsymbol{\vartheta},\boldsymbol{w}) = w_j^{(i)}w_{j'}^{(i)}\,\mathbb{E}_{\Pi,\boldsymbol{x}\sim\mathcal{D}(\mathcal{X})}\left[\partial^2_{\sigma^{(i)}\sigma^{(i)}}\ell\left(\widetilde{\sigma}_\Pi(\boldsymbol{x};\boldsymbol{\vartheta}^{(1)}),\cdots,\widetilde{\sigma}_\Pi(\boldsymbol{x};\boldsymbol{\vartheta}^{(n)})\right)\nabla\sigma_\Pi(\boldsymbol{x};\boldsymbol{\vartheta}_j^{(i)})\nabla\sigma_\Pi(\boldsymbol{x};\boldsymbol{\vartheta}_{j'}^{(i)})^\top\right]\tag{23}$$

3. When $i \neq i'$, $\partial_{\boldsymbol{\vartheta}_{j'}^{(i')}}\nabla\sigma_\Pi(\boldsymbol{x};\boldsymbol{\vartheta}_j^{(i)}) = \mathbf{0}$, and henceforth

$$\partial^2_{\boldsymbol{\vartheta}_j^{(i)}\boldsymbol{\vartheta}_{j'}^{(i')}}\mathcal{L}(\boldsymbol{\vartheta},\boldsymbol{w}) = w_j^{(i)}w_{j'}^{(i')}\,\mathbb{E}_{\Pi,\boldsymbol{x}\sim\mathcal{D}(\mathcal{X})}\left[\partial^2_{\sigma^{(i)}\sigma^{(i')}}\ell\left(\widetilde{\sigma}_\Pi(\boldsymbol{x};\boldsymbol{\vartheta}^{(1)}),\cdots,\widetilde{\sigma}_\Pi(\boldsymbol{x};\boldsymbol{\vartheta}^{(n)})\right)\nabla\sigma_\Pi(\boldsymbol{x};\boldsymbol{\vartheta}_j^{(i)})\nabla\sigma_\Pi(\boldsymbol{x};\boldsymbol{\vartheta}_{j'}^{(i')})^\top\right]\tag{24}$$

Using this fact again: $\boldsymbol{\vartheta}_j^{(i)} = \boldsymbol{\theta}^{(i)}$ and $\widetilde{\sigma}_\Pi(\boldsymbol{x};\boldsymbol{\vartheta}^{(i)}) = \sum_{j=1}^{m_i} w_j^{(i)}\sigma_\Pi(\boldsymbol{x};\boldsymbol{\theta}^{(i)}) = \sigma_\Pi(\boldsymbol{x};\boldsymbol{\theta}^{(i)})$ when $\epsilon = 0$, Eq. 22 becomes:

$$\begin{aligned}
\partial^2_{\boldsymbol{\vartheta}_j^{(i)}\boldsymbol{\vartheta}_j^{(i)}}\mathcal{L}(\boldsymbol{\vartheta},\boldsymbol{w})\Big|_{\epsilon=0} &= w_j^{(i)2}\,\mathbb{E}_{\Pi,\boldsymbol{x}\sim\mathcal{D}(\mathcal{X})}\left[\partial^2_{\sigma^{(i)}\sigma^{(i)}}\ell\left(\sigma_\Pi(\boldsymbol{x};\boldsymbol{\theta}^{(1)}),\cdots,\sigma_\Pi(\boldsymbol{x};\boldsymbol{\theta}^{(n)})\right)\nabla\sigma_\Pi(\boldsymbol{x};\boldsymbol{\theta}^{(i)})\nabla\sigma_\Pi(\boldsymbol{x};\boldsymbol{\theta}^{(i)})^\top\right]\\
&\quad + w_j^{(i)}\,\mathbb{E}_{\Pi,\boldsymbol{x}\sim\mathcal{D}(\mathcal{X})}\left[\partial_{\sigma^{(i)}}\ell\left(\sigma_\Pi(\boldsymbol{x};\boldsymbol{\theta}^{(1)}),\cdots,\sigma_\Pi(\boldsymbol{x};\boldsymbol{\theta}^{(n)})\right)\nabla^2\sigma_\Pi(\boldsymbol{x};\boldsymbol{\theta}^{(i)})\right]\\
&= w_j^{(i)2}\boldsymbol{T}^{(i)}(\boldsymbol{\theta}) + w_j^{(i)}\boldsymbol{S}^{(i)}(\boldsymbol{\theta}),
\end{aligned}$$

Eq. 23 can be simplified as:

$$\begin{aligned}
\partial^2_{\boldsymbol{\vartheta}_j^{(i)}\boldsymbol{\vartheta}_{j'}^{(i)}}\mathcal{L}(\boldsymbol{\vartheta},\boldsymbol{w})\Big|_{\epsilon=0} &= w_j^{(i)}w_{j'}^{(i)}\,\mathbb{E}_{\Pi,\boldsymbol{x}\sim\mathcal{D}(\mathcal{X})}\left[\partial^2_{\sigma^{(i)}\sigma^{(i)}}\ell\left(\sigma_\Pi(\boldsymbol{x};\boldsymbol{\theta}^{(1)}),\cdots,\sigma_\Pi(\boldsymbol{x};\boldsymbol{\theta}^{(n)})\right)\nabla\sigma_\Pi(\boldsymbol{x};\boldsymbol{\theta}^{(i)})\nabla\sigma_\Pi(\boldsymbol{x};\boldsymbol{\theta}^{(i)})^\top\right]\\
&= w_j^{(i)}w_{j'}^{(i)}\boldsymbol{T}^{(i)}(\boldsymbol{\theta}),
\end{aligned}$$

and by Lemma 3, Eq. 24 turns into:

$$\begin{aligned}
\partial^2_{\boldsymbol{\vartheta}_j^{(i)}\boldsymbol{\vartheta}_{j'}^{(i')}}\mathcal{L}(\boldsymbol{\vartheta},\boldsymbol{w})\Big|_{\epsilon=0} &= w_j^{(i)}w_{j'}^{(i')}\,\mathbb{E}_{\Pi,\boldsymbol{x}\sim\mathcal{D}(\mathcal{X})}\left[\partial^2_{\sigma^{(i)}\sigma^{(i')}}\ell\left(\sigma_\Pi(\boldsymbol{x};\boldsymbol{\theta}^{(1)}),\cdots,\sigma_\Pi(\boldsymbol{x};\boldsymbol{\theta}^{(n)})\right)\nabla\sigma_\Pi(\boldsymbol{x};\boldsymbol{\theta}^{(i)})\nabla\sigma_\Pi(\boldsymbol{x};\boldsymbol{\theta}^{(i')})^\top\right]\\
&= w_j^{(i)}w_{j'}^{(i')}\partial^2_{\boldsymbol{\vartheta}_j^{(i)}\boldsymbol{\vartheta}_{j'}^{(i')}}\mathcal{L}(\boldsymbol{\theta}),
\end{aligned}$$

all as desired. □

Lemma 5. *The following equalities hold for $\mathcal{L}(\boldsymbol{\theta}^{(\setminus i)}, \boldsymbol{\vartheta}^{(i)}, \boldsymbol{w}^{(i)})$ at $\epsilon = 0$ for any $i \in [n]$:*

$$\partial_{\boldsymbol{\vartheta}_j^{(i)}} \mathcal{L}(\boldsymbol{\theta}^{(\setminus i)}, \boldsymbol{\vartheta}^{(i)}, \boldsymbol{w}^{(i)})\Big|_{\epsilon=0} = w_j^{(i)} \partial_{\boldsymbol{\theta}^{(i)}} \mathcal{L}(\boldsymbol{\theta}), \quad \forall j \in [m_i], \tag{25}$$

$$\partial^2_{\boldsymbol{\vartheta}_j^{(i)} \boldsymbol{\vartheta}_j^{(i)}} \mathcal{L}(\boldsymbol{\theta}^{(\setminus i)}, \boldsymbol{\vartheta}^{(i)}, \boldsymbol{w}^{(i)})\Big|_{\epsilon=0} = w_j^{(i)} \boldsymbol{S}^{(i)}(\boldsymbol{\theta}) + w_j^{(i)2} \boldsymbol{T}^{(i)}(\boldsymbol{\theta}), \quad \forall j \in [m_i], \tag{26}$$

$$\partial^2_{\boldsymbol{\vartheta}_j^{(i)} \boldsymbol{\vartheta}_{j'}^{(i)}} \mathcal{L}(\boldsymbol{\theta}^{(\setminus i)}, \boldsymbol{\vartheta}^{(i)}, \boldsymbol{w}^{(i)})\Big|_{\epsilon=0} = w_j^{(i)} w_{j'}^{(i)} \boldsymbol{T}^{(i)}(\boldsymbol{\theta}), \quad \forall j, j' \in [m_i], j \neq j', \tag{27}$$

where $\boldsymbol{T}^{(i)}(\boldsymbol{\theta}) = \mathbb{E}_{\Pi, \boldsymbol{x} \sim \mathcal{D}(\mathcal{X})} \left[\partial^2_{\sigma^{(i)} \sigma^{(i)}} \ell \left(\sigma_\Pi(\boldsymbol{x}; \boldsymbol{\theta}^{(1)}), \cdots, \sigma_\Pi(\boldsymbol{x}; \boldsymbol{\theta}^{(n)}) \right) \nabla \sigma_\Pi(\boldsymbol{x}; \boldsymbol{\theta}^{(i)}) \nabla \sigma_\Pi(\boldsymbol{x}; \boldsymbol{\theta}^{(i)})^\top \right]$ is as defined in Lemma 3.

Proof. The proof is identical to Lemma 4. We outline the details for completeness. Let $\widetilde{\sigma}_\Pi(\boldsymbol{x}; \boldsymbol{\vartheta}^{(i)}) = \sum_{j=1}^{m_i} w_j^{(i)} \sigma_\Pi(\boldsymbol{x}; \boldsymbol{\vartheta}_j^{(i)})$ and we can express $\mathcal{L}(\boldsymbol{\theta}^{(\setminus i)}, \boldsymbol{\vartheta}^{(i)}, \boldsymbol{w}^{(i)})$ as:

$$\mathcal{L}(\boldsymbol{\theta}^{(\setminus i)}, \boldsymbol{\vartheta}^{(i)}, \boldsymbol{w}^{(i)}) = \mathbb{E}_{\Pi, \boldsymbol{x} \sim \mathcal{D}(\mathcal{X})} \left[\ell \left(\sigma_\Pi(\boldsymbol{x}; \boldsymbol{\theta}^{(1)}), \cdots, \widetilde{\sigma}_\Pi(\boldsymbol{x}; \boldsymbol{\vartheta}^{(i)}), \cdots, \sigma_\Pi(\boldsymbol{x}; \boldsymbol{\theta}^{(n)}) \right) \right]. \tag{28}$$

By chain rule, for every $j \in [m_i]$:

$$\begin{aligned}
&\partial_{\boldsymbol{\vartheta}_j^{(i)}} \mathcal{L}(\boldsymbol{\theta}^{(\setminus i)}, \boldsymbol{\vartheta}^{(i)}, \boldsymbol{w}^{(i)}) = \partial_{\boldsymbol{\vartheta}_j^{(i)}} \mathbb{E}_{\Pi, \boldsymbol{x} \sim \mathcal{D}(\mathcal{X})} \left[\ell \left(\sigma_\Pi(\boldsymbol{x}; \boldsymbol{\theta}^{(1)}), \cdots, \widetilde{\sigma}_\Pi(\boldsymbol{x}; \boldsymbol{\vartheta}^{(i)}), \cdots, \sigma_\Pi(\boldsymbol{x}; \boldsymbol{\theta}^{(n)}) \right) \right] \\
&= \mathbb{E}_{\Pi, \boldsymbol{x} \sim \mathcal{D}(\mathcal{X})} \left[\partial_{\sigma^{(i)}} \ell \left(\sigma_\Pi(\boldsymbol{x}; \boldsymbol{\theta}^{(1)}), \cdots, \widetilde{\sigma}_\Pi(\boldsymbol{x}; \boldsymbol{\vartheta}^{(i)}), \cdots, \sigma_\Pi(\boldsymbol{x}; \boldsymbol{\theta}^{(n)}) \right) \partial_{\boldsymbol{\vartheta}_j^{(i)}} \widetilde{\sigma}_\Pi(\boldsymbol{x}; \boldsymbol{\vartheta}^{(i)}) \right] \\
&= w_j^{(i)} \mathbb{E}_{\Pi, \boldsymbol{x} \sim \mathcal{D}(\mathcal{X})} \left[\partial_{\sigma^{(i)}} \ell \left(\sigma_\Pi(\boldsymbol{x}; \boldsymbol{\theta}^{(1)}), \cdots, \widetilde{\sigma}_\Pi(\boldsymbol{x}; \boldsymbol{\vartheta}^{(i)}), \cdots, \sigma_\Pi(\boldsymbol{x}; \boldsymbol{\theta}^{(n)}) \right) \nabla \sigma_\Pi(\boldsymbol{x}; \boldsymbol{\vartheta}_j^{(i)}) \right].
\end{aligned} \tag{29}$$

Using the fact that $\boldsymbol{\vartheta}_j^{(i)} = \boldsymbol{\theta}^{(i)}$ and $\widetilde{\sigma}_\Pi(\boldsymbol{x}; \boldsymbol{\vartheta}^{(i)}) = \sum_{j=1}^{m_i} w_j^{(i)} \sigma_\Pi(\boldsymbol{x}; \boldsymbol{\theta}^{(i)}) = \sigma_\Pi(\boldsymbol{x}; \boldsymbol{\theta}^{(i)})$ when $\epsilon = 0$, Eq. 29 can be rewritten as:

$$\begin{aligned}
\partial_{\boldsymbol{\vartheta}_j^{(i)}} \mathcal{L}(\boldsymbol{\theta}^{(\setminus i)}, \boldsymbol{\vartheta}^{(i)}, \boldsymbol{w}^{(i)})\Big|_{\epsilon=0} &= w_j^{(i)} \mathbb{E}_{\Pi, \boldsymbol{x} \sim \mathcal{D}(\mathcal{X})} \left[\partial_{\sigma^{(i)}} \ell \left(\sigma_\Pi(\boldsymbol{x}; \boldsymbol{\theta}^{(1)}), \cdots, \widetilde{\sigma}_\Pi(\boldsymbol{x}; \boldsymbol{\vartheta}^{(i)}), \cdots, \sigma_\Pi(\boldsymbol{x}; \boldsymbol{\theta}^{(n)}) \right) \nabla \sigma_\Pi(\boldsymbol{x}; \boldsymbol{\vartheta}_j^{(i)}) \right] \Big|_{\epsilon=0} \\
&= w_j^{(i)} \mathbb{E}_{\Pi, \boldsymbol{x} \sim \mathcal{D}(\mathcal{X})} \left[\partial_{\sigma^{(i)}} \ell \left(\sigma_\Pi(\boldsymbol{x}; \boldsymbol{\theta}^{(1)}), \cdots, \sigma_\Pi(\boldsymbol{x}; \boldsymbol{\theta}^{(i)}), \cdots, \sigma_\Pi(\boldsymbol{x}; \boldsymbol{\theta}^{(n)}) \right) \nabla \sigma_\Pi(\boldsymbol{x}; \boldsymbol{\theta}^{(i)}) \right] \\
&= w_j^{(i)} \partial_{\boldsymbol{\theta}^{(i)}} \mathcal{L}(\boldsymbol{\theta}),
\end{aligned}$$

where the last step is due to Lemma 3.

Next we derive second-order derivatives. Taking derivatives of Eq. 29 in terms of $\boldsymbol{\vartheta}_{j'}^{(i)}$ for some $j' \in [m_i]$, and by chain rule:

$$\begin{aligned}
&\partial^2_{\boldsymbol{\vartheta}_j^{(i)} \boldsymbol{\vartheta}_{j'}^{(i)}} \mathcal{L}(\boldsymbol{\theta}^{(\setminus i)}, \boldsymbol{\vartheta}^{(i)}, \boldsymbol{w}^{(i)}) = \partial_{\boldsymbol{\vartheta}_{j'}^{(i)}} w_j^{(i)} \mathbb{E}_{\Pi, \boldsymbol{x} \sim \mathcal{D}(\mathcal{X})} \left[\partial_{\sigma^{(i)}} \ell \left(\cdots, \widetilde{\sigma}_\Pi(\boldsymbol{x}; \boldsymbol{\vartheta}^{(i)}), \cdots \right) \nabla \sigma_\Pi(\boldsymbol{x}; \boldsymbol{\vartheta}_j^{(i)}) \right] \\
&= w_j^{(i)} \mathbb{E}_{\Pi, \boldsymbol{x} \sim \mathcal{D}(\mathcal{X})} \left[\partial^2_{\sigma^{(i)} \sigma^{(i)}} \ell \left(\cdots, \widetilde{\sigma}_\Pi(\boldsymbol{x}; \boldsymbol{\vartheta}^{(i)}), \cdots \right) \nabla \sigma_\Pi(\boldsymbol{x}; \boldsymbol{\vartheta}_j^{(i)}) \partial_{\boldsymbol{\vartheta}_{j'}^{(i)}} \widetilde{\sigma}_\Pi(\boldsymbol{x}; \boldsymbol{\vartheta}^{(i')})^\top \right] \\
&\quad + w_j^{(i)} \mathbb{E}_{\Pi, \boldsymbol{x} \sim \mathcal{D}(\mathcal{X})} \left[\partial_{\sigma^{(i)}} \ell \left(\cdots, \widetilde{\sigma}_\Pi(\boldsymbol{x}; \boldsymbol{\vartheta}^{(i)}), \cdots \right) \partial_{\boldsymbol{\vartheta}_{j'}^{(i)}} \nabla \sigma_\Pi(\boldsymbol{x}; \boldsymbol{\vartheta}_j^{(i)}) \right] \\
&= w_j^{(i)} w_{j'}^{(i)} \mathbb{E}_{\Pi, \boldsymbol{x} \sim \mathcal{D}(\mathcal{X})} \left[\partial^2_{\sigma^{(i)} \sigma^{(i)}} \ell \left(\cdots, \widetilde{\sigma}_\Pi(\boldsymbol{x}; \boldsymbol{\vartheta}^{(i)}), \cdots \right) \nabla \sigma_\Pi(\boldsymbol{x}; \boldsymbol{\vartheta}_j^{(i)}) \nabla \sigma_\Pi(\boldsymbol{x}; \boldsymbol{\vartheta}_{j'}^{(i)})^\top \right] \\
&\quad + w_j^{(i)} \mathbb{E}_{\Pi, \boldsymbol{x} \sim \mathcal{D}(\mathcal{X})} \left[\partial_{\sigma^{(i)}} \ell \left(\cdots, \widetilde{\sigma}_\Pi(\boldsymbol{x}; \boldsymbol{\vartheta}^{(i)}), \cdots \right) \partial_{\boldsymbol{\vartheta}_{j'}^{(i)}} \nabla \sigma_\Pi(\boldsymbol{x}; \boldsymbol{\vartheta}_j^{(i)}) \right].
\end{aligned}$$

Now we consider two scenarios:
1. When $j = j'$, $\partial_{\boldsymbol{\vartheta}_{j'}^{(i)}} \nabla \sigma_\Pi(\boldsymbol{x}; \boldsymbol{\vartheta}_j^{(i)}) = \nabla^2 \sigma_\Pi(\boldsymbol{x}; \boldsymbol{\vartheta}_j^{(i)})$, and then

$$\begin{aligned}
\partial^2_{\boldsymbol{\vartheta}_j^{(i)} \boldsymbol{\vartheta}_j^{(i)}} \mathcal{L}(\boldsymbol{\theta}^{(\setminus i)}, \boldsymbol{\vartheta}^{(i)}, \boldsymbol{w}^{(i)})\Big|_{\epsilon=0} &= w_j^{(i)2} \mathbb{E}_{\Pi, \boldsymbol{x} \sim \mathcal{D}(\mathcal{X})} \left[\partial^2_{\sigma^{(i)} \sigma^{(i)}} \ell \left(\cdots, \widetilde{\sigma}_\Pi(\boldsymbol{x}; \boldsymbol{\vartheta}^{(i)}), \cdots \right) \nabla \sigma_\Pi(\boldsymbol{x}; \boldsymbol{\vartheta}_j^{(i)}) \nabla \sigma_\Pi(\boldsymbol{x}; \boldsymbol{\vartheta}_j^{(i)})^\top \right] \\
&\quad + w_j^{(i)} \mathbb{E}_{\Pi, \boldsymbol{x} \sim \mathcal{D}(\mathcal{X})} \left[\partial_{\sigma^{(i)}} \ell \left(\cdots, \widetilde{\sigma}_\Pi(\boldsymbol{x}; \boldsymbol{\vartheta}^{(i)}), \cdots \right) \nabla^2 \sigma_\Pi(\boldsymbol{x}; \boldsymbol{\vartheta}_j^{(i)}) \right]
\end{aligned} \tag{30}$$

2. When $j \neq j'$, $\partial_{\boldsymbol{\vartheta}_{j'}^{(i)}} \nabla \sigma_\Pi(\boldsymbol{x}; \boldsymbol{\vartheta}_j^{(i)}) = \mathbf{0}$, and thus

$$\partial^2_{\boldsymbol{\vartheta}_j^{(i)} \boldsymbol{\vartheta}_{j'}^{(i)}} \mathcal{L}(\boldsymbol{\theta}^{(\setminus i)}, \boldsymbol{\vartheta}^{(i)}, \boldsymbol{w}^{(i)}) \Big|_{\epsilon=0} = w_j^{(i)} w_{j'}^{(i)} \, \mathbb{E}_{\Pi, \boldsymbol{x} \sim \mathcal{D}(\mathcal{X})} \left[\partial^2_{\sigma^{(i)} \sigma^{(i)}} \ell \left(\cdots, \widetilde{\sigma}_\Pi(\boldsymbol{x}; \boldsymbol{\vartheta}^{(i)}), \cdots \right) \nabla \sigma_\Pi(\boldsymbol{x}; \boldsymbol{\vartheta}_j^{(i)}) \nabla \sigma_\Pi(\boldsymbol{x}; \boldsymbol{\vartheta}_{j'}^{(i)})^\top \right] \tag{31}$$

Using this fact again: $\boldsymbol{\vartheta}_j^{(i)} = \boldsymbol{\theta}^{(i)}$ and $\widetilde{\sigma}_\Pi(\boldsymbol{x}; \boldsymbol{\vartheta}^{(i)}) = \sum_{j=1}^{m_i} w_j^{(i)} \sigma_\Pi(\boldsymbol{x}; \boldsymbol{\theta}^{(i)}) = \sigma_\Pi(\boldsymbol{x}; \boldsymbol{\theta}^{(i)})$ when $\epsilon = 0$, Eq. 30 becomes:

$$\begin{aligned}
\partial^2_{\boldsymbol{\vartheta}_j^{(i)} \boldsymbol{\vartheta}_j^{(i)}} \mathcal{L}(\boldsymbol{\theta}^{(\setminus i)}, \boldsymbol{\vartheta}^{(i)}, \boldsymbol{w}^{(i)}) &= w_j^{(i)2} \, \mathbb{E}_{\Pi, \boldsymbol{x} \sim \mathcal{D}(\mathcal{X})} \left[\partial^2_{\sigma^{(i)} \sigma^{(i)}} \ell \left(\cdots, \sigma_\Pi(\boldsymbol{x}; \boldsymbol{\theta}^{(i)}), \cdots \right) \nabla \sigma_\Pi(\boldsymbol{x}; \boldsymbol{\theta}^{(i)}) \nabla \sigma_\Pi(\boldsymbol{x}; \boldsymbol{\theta}^{(i)})^\top \right] \\
&\quad + w_j^{(i)} \, \mathbb{E}_{\Pi, \boldsymbol{x} \sim \mathcal{D}(\mathcal{X})} \left[\partial_{\sigma^{(i)}} \ell \left(\cdots, \sigma_\Pi(\boldsymbol{x}; \boldsymbol{\theta}^{(i)}), \cdots \right) \nabla^2 \sigma_\Pi(\boldsymbol{x}; \boldsymbol{\theta}^{(i)}) \right] \\
&= w_j^{(i)2} \boldsymbol{T}^{(i)}(\boldsymbol{\theta}) + w_j^{(i)} \boldsymbol{S}^{(i)}(\boldsymbol{\theta}),
\end{aligned}$$

and Eq. 31 can be simplified as:

$$\begin{aligned}
\partial^2_{\boldsymbol{\vartheta}_j^{(i)} \boldsymbol{\vartheta}_{j'}^{(i)}} \mathcal{L}(\boldsymbol{\theta}^{(\setminus i)}, \boldsymbol{\vartheta}^{(i)}, \boldsymbol{w}^{(i)}) &= w_j^{(i)} w_{j'}^{(i)} \, \mathbb{E}_{\Pi, \boldsymbol{x} \sim \mathcal{D}(\mathcal{X})} \left[\partial^2_{\sigma^{(i)} \sigma^{(i)}} \ell \left(\cdots, \sigma_\Pi(\boldsymbol{x}; \boldsymbol{\theta}^{(i)}), \cdots \right) \nabla \sigma_\Pi(\boldsymbol{x}; \boldsymbol{\theta}^{(i)}) \nabla \sigma_\Pi(\boldsymbol{x}; \boldsymbol{\theta}^{(i)})^\top \right] \\
&= w_j^{(i)} w_{j'}^{(i)} \boldsymbol{T}^{(i)}(\boldsymbol{\theta}),
\end{aligned}$$

both as desired. □

Lemma 6. *Assume $\mathcal{L}(\boldsymbol{\vartheta}, \boldsymbol{w})$ has bounded third-order derivatives with respect to $\boldsymbol{\vartheta}$, then we have*

$$(\mathcal{L}(\boldsymbol{\vartheta}, \boldsymbol{w}) - \mathcal{L}(\boldsymbol{\theta})) = \sum_{i=1}^n \left(\mathcal{L}(\boldsymbol{\theta}^{(\setminus i)}, \boldsymbol{\vartheta}^{(i)}, \boldsymbol{w}^{(i)}) - \mathcal{L}(\boldsymbol{\theta}) \right) + \frac{\epsilon^2}{2} \sum_{\substack{i, i' \in [n] \\ i \neq i'}} \boldsymbol{\mu}^{(i)\top} \partial^2_{\boldsymbol{\theta}^{(i)} \boldsymbol{\theta}^{(i')}} \mathcal{L}(\boldsymbol{\theta}) \boldsymbol{\mu}^{(i')} + \mathcal{O}(\epsilon^3),$$

where $\mathcal{L}(\boldsymbol{\theta}^{(\setminus i)}, \boldsymbol{\vartheta}^{(i)}, \boldsymbol{w}^{(i)})$ and $\boldsymbol{\mu}^{(i)}$ are as defined in Theorem 1.

Proof. Define an auxiliary function:

$$F(\epsilon) = (\mathcal{L}(\boldsymbol{\vartheta}, \boldsymbol{w}) - \mathcal{L}(\boldsymbol{\theta})) - \sum_{i=1}^n \left(\mathcal{L}(\boldsymbol{\theta}^{(\setminus i)}, \boldsymbol{\vartheta}^{(i)}, \boldsymbol{w}^{(i)}) - \mathcal{L}(\boldsymbol{\theta}) \right).$$

Note that $F(\epsilon)$ also has bounded third-order derivatives. Hence, by Taylor expansion:

$$F(\epsilon) = F(0) + \epsilon \frac{d}{d\epsilon} F(0) + \frac{\epsilon^2}{2} \frac{d^2}{d\epsilon^2} F(0) + \mathcal{O}(\epsilon^3). \tag{32}$$

Compute the first-order derivatives of F via path derivatives, we can derive

$$\frac{d}{d\epsilon} (\mathcal{L}(\boldsymbol{\vartheta}, \boldsymbol{w}) - \mathcal{L}(\boldsymbol{\theta})) = \sum_{i=1}^n \sum_{j=1}^{m_i} \partial_{\boldsymbol{\vartheta}_j^{(i)}} \mathcal{L}(\boldsymbol{\vartheta}, \boldsymbol{w})^\top \frac{d \boldsymbol{\vartheta}_j^{(i)}}{d\epsilon} = \sum_{i=1}^n \sum_{j=1}^{m_i} \partial_{\boldsymbol{\vartheta}_j^{(i)}} \mathcal{L}(\boldsymbol{\vartheta}, \boldsymbol{w})^\top \boldsymbol{\Delta}_j^{(i)}, \tag{33}$$

and for every $i \in [n]$:

$$\begin{aligned}
\frac{d}{d\epsilon} \left(\mathcal{L}(\boldsymbol{\theta}^{(\setminus i)}, \boldsymbol{\vartheta}^{(i)}, \boldsymbol{w}^{(i)}) - \mathcal{L}(\boldsymbol{\theta}) \right) &= \sum_{j=1}^{m_i} \partial_{\boldsymbol{\vartheta}_j^{(i)}} \mathcal{L}(\boldsymbol{\theta}^{(\setminus i)}, \boldsymbol{\vartheta}^{(i)}, \boldsymbol{w}^{(i)})^\top \frac{d \boldsymbol{\vartheta}_j^{(i)}}{d\epsilon} \\
&= \sum_{j=1}^{m_i} \partial_{\boldsymbol{\vartheta}_j^{(i)}} \mathcal{L}(\boldsymbol{\theta}^{(\setminus i)}, \boldsymbol{\vartheta}^{(i)}, \boldsymbol{w}^{(i)})^\top \boldsymbol{\Delta}_j^{(i)},
\end{aligned} \tag{34}$$

By Lemma 4 and Lemma 5, $\partial_{\boldsymbol{\vartheta}_j^{(i)}} \mathcal{L}(\boldsymbol{\theta}^{(\setminus i)}, \boldsymbol{\vartheta}^{(i)}, \boldsymbol{w}^{(i)})\Big|_{\epsilon=0} = \partial_{\boldsymbol{\vartheta}_j^{(i)}} \mathcal{L}(\boldsymbol{\vartheta}, \boldsymbol{w})\Big|_{\epsilon=0}$, hence combining Eq. 33 and 34:

$$\frac{d}{d\epsilon} F(0) = \left[\sum_{i=1}^{n} \sum_{j=1}^{m_i} \partial_{\boldsymbol{\vartheta}_j^{(i)}} \mathcal{L}(\boldsymbol{\vartheta}, \boldsymbol{w})^\top \boldsymbol{\Delta}_j^{(i)} - \sum_{i=1}^{n} \sum_{j=1}^{m_i} \partial_{\boldsymbol{\vartheta}_j^{(i)}} \mathcal{L}(\boldsymbol{\theta}^{(\setminus i)}, \boldsymbol{\vartheta}^{(i)}, \boldsymbol{w}^{(i)})^\top \boldsymbol{\Delta}_j^{(i)} \right] \Bigg|_{\epsilon=0} = 0. \tag{35}$$

We can also compute the second-order derivatives via path derivatives:

$$\begin{aligned}
\frac{d^2}{d\epsilon^2} \left(\mathcal{L}(\boldsymbol{\vartheta}, \boldsymbol{w}) - \mathcal{L}(\boldsymbol{\theta}) \right) &= \frac{d}{d\epsilon} \sum_{i=1}^{n} \sum_{j=1}^{m_i} \partial_{\boldsymbol{\vartheta}_j^{(i)}} \mathcal{L}(\boldsymbol{\vartheta}, \boldsymbol{w})^\top \boldsymbol{\Delta}_j^{(i)} \\
&= \sum_{i=1}^{n} \sum_{i'=1}^{n} \sum_{j=1}^{m_i} \sum_{j'=1}^{m_{i'}} \boldsymbol{\Delta}_j^{(i)\top} \partial^2_{\boldsymbol{\vartheta}_j^{(i)} \boldsymbol{\vartheta}_{j'}^{(i')}} \mathcal{L}(\boldsymbol{\vartheta}, \boldsymbol{w}) \frac{d\boldsymbol{\vartheta}_{j'}^{(i')}}{d\epsilon} \\
&= \sum_{i=1}^{n} \sum_{i'=1}^{n} \sum_{j=1}^{m_i} \sum_{j'=1}^{m_{i'}} \boldsymbol{\Delta}_j^{(i)\top} \partial^2_{\boldsymbol{\vartheta}_j^{(i)} \boldsymbol{\vartheta}_{j'}^{(i')}} \mathcal{L}(\boldsymbol{\vartheta}, \boldsymbol{w}) \boldsymbol{\Delta}_{j'}^{(i')},
\end{aligned} \tag{36}$$

and similarly for every $i \in [n]$,

$$\begin{aligned}
\frac{d^2}{d\epsilon^2} \left(\mathcal{L}(\boldsymbol{\theta}^{(\setminus i)}, \boldsymbol{\vartheta}^{(i)}, \boldsymbol{w}^{(i)}) - \mathcal{L}(\boldsymbol{\theta}) \right) &= \frac{d}{d\epsilon} \sum_{j=1}^{m_i} \partial_{\boldsymbol{\vartheta}_j^{(i)}} \mathcal{L}(\boldsymbol{\theta}^{(\setminus i)}, \boldsymbol{\vartheta}^{(i)}, \boldsymbol{w}^{(i)})^\top \boldsymbol{\Delta}_j^{(i)} \\
&= \sum_{j=1}^{m_i} \sum_{j'=1}^{m_i} \boldsymbol{\Delta}_j^{(i)\top} \partial^2_{\boldsymbol{\vartheta}_j^{(i)} \boldsymbol{\vartheta}_{j'}^{(i)}} \mathcal{L}(\boldsymbol{\theta}^{(\setminus i)}, \boldsymbol{\vartheta}^{(i)}, \boldsymbol{w}^{(i)}) \frac{d\boldsymbol{\vartheta}_{j'}^{(i)}}{d\epsilon} \\
&= \sum_{j=1}^{m_i} \sum_{j'=1}^{m_i} \boldsymbol{\Delta}_j^{(i)\top} \partial^2_{\boldsymbol{\vartheta}_j^{(i)} \boldsymbol{\vartheta}_{j'}^{(i)}} \mathcal{L}(\boldsymbol{\theta}^{(\setminus i)}, \boldsymbol{\vartheta}^{(i)}, \boldsymbol{w}^{(i)}) \boldsymbol{\Delta}_{j'}^{(i)}.
\end{aligned} \tag{37}$$

By Lemma 4 and Lemma 5, $\partial^2_{\boldsymbol{\vartheta}_j^{(i)} \boldsymbol{\vartheta}_{j'}^{(i)}} \mathcal{L}(\boldsymbol{\theta}^{(\setminus i)}, \boldsymbol{\vartheta}^{(i)}, \boldsymbol{w}^{(i)})\Big|_{\epsilon=0} = \partial^2_{\boldsymbol{\vartheta}_j^{(i)} \boldsymbol{\vartheta}_{j'}^{(i)}} \mathcal{L}(\boldsymbol{\vartheta}, \boldsymbol{w})\Big|_{\epsilon=0}$ for any $i \in [n]$ and $j, j' \in [m_i]$, hence we can cancel all terms in Eq. 37 by:

$$\begin{aligned}
\frac{d^2}{d\epsilon^2} F(0) &= \left[\sum_{i,i' \in [n]} \sum_{j=1}^{m_i} \sum_{j'=1}^{m_{i'}} \boldsymbol{\Delta}_j^{(i)\top} \partial^2_{\boldsymbol{\vartheta}_j^{(i)} \boldsymbol{\vartheta}_{j'}^{(i')}} \mathcal{L}(\boldsymbol{\vartheta}, \boldsymbol{w}) \boldsymbol{\Delta}_{j'}^{(i')} - \sum_{i=1}^{n} \sum_{j=1}^{m_i} \sum_{j'=1}^{m_i} \boldsymbol{\Delta}_j^{(i)\top} \partial^2_{\boldsymbol{\vartheta}_j^{(i)} \boldsymbol{\vartheta}_{j'}^{(i)}} \mathcal{L}(\boldsymbol{\theta}^{(\setminus i)}, \boldsymbol{\vartheta}^{(i)}, \boldsymbol{w}^{(i)}) \boldsymbol{\Delta}_{j'}^{(i)} \right] \Bigg|_{\epsilon=0} \\
&= \left[\sum_{i=1}^{n} \sum_{j=1}^{m_i} \sum_{j'=1}^{m_i} \boldsymbol{\Delta}_j^{(i)\top} \left(\partial^2_{\boldsymbol{\vartheta}_j^{(i)} \boldsymbol{\vartheta}_{j'}^{(i)}} \mathcal{L}(\boldsymbol{\vartheta}, \boldsymbol{w}) - \partial^2_{\boldsymbol{\vartheta}_j^{(i)} \boldsymbol{\vartheta}_{j'}^{(i)}} \mathcal{L}(\boldsymbol{\theta}^{(\setminus i)}, \boldsymbol{\vartheta}^{(i)}, \boldsymbol{w}^{(i)}) \right) \boldsymbol{\Delta}_{j'}^{(i)} \right] \Bigg|_{\epsilon=0} \\
&\quad + \left[\sum_{i \neq i'} \sum_{j=1}^{m_i} \sum_{j'=1}^{m_{i'}} \boldsymbol{\Delta}_j^{(i)\top} \partial^2_{\boldsymbol{\vartheta}_j^{(i)} \boldsymbol{\vartheta}_{j'}^{(i')}} \mathcal{L}(\boldsymbol{\vartheta}, \boldsymbol{w}) \boldsymbol{\Delta}_{j'}^{(i')} \right] \Bigg|_{\epsilon=0} \\
&= \sum_{i \neq i'} \sum_{j=1}^{m_i} \sum_{j'=1}^{m_{i'}} w_j^{(i)} w_{j'}^{(i')} \boldsymbol{\Delta}_j^{(i)\top} \partial^2_{\boldsymbol{\vartheta}_j^{(i)} \boldsymbol{\vartheta}_{j'}^{(i')}} \mathcal{L}(\boldsymbol{\theta}) \boldsymbol{\Delta}_{j'}^{(i')} \\
&= \sum_{i \neq i'} \sum_{j=1}^{m_i} \sum_{j'=1}^{m_{i'}} \boldsymbol{\mu}^{(i)\top} \partial^2_{\boldsymbol{\vartheta}_j^{(i)} \boldsymbol{\vartheta}_{j'}^{(i')}} \mathcal{L}(\boldsymbol{\theta}) \boldsymbol{\mu}^{(i')},
\end{aligned} \tag{38}$$

where we use Eq. 19 in Lemma 4 for the last second equality, and we use the fact: $\sum_{j=1}^{m_i} w_j^{(i)} \boldsymbol{\Delta}_j^{(i)} = \boldsymbol{\mu}^{(i)}$ to get the last equality. Merging Eq. 32, 35, 38, we obtain the result as desired. □

Lemma 7. *Assume $\mathcal{L}(\boldsymbol{\vartheta}, \boldsymbol{w})$ has bounded third-order derivatives with respect to $\boldsymbol{\vartheta}$, then we have*

$$\mathcal{L}(\boldsymbol{\theta}^{(\backslash i)}, \boldsymbol{\vartheta}^{(i)}, \boldsymbol{w}^{(i)}) - \mathcal{L}(\boldsymbol{\theta}) = \epsilon \partial_{\boldsymbol{\theta}^{(i)}} \mathcal{L}(\boldsymbol{\theta})^\top \boldsymbol{\mu}^{(i)} + \frac{\epsilon^2}{2} \boldsymbol{\mu}^{(i)\top} \partial^2_{\boldsymbol{\theta}^{(i)}\boldsymbol{\theta}^{(i)}} \mathcal{L}(\boldsymbol{\theta}) \boldsymbol{\mu}^{(i)} + \frac{\epsilon^2}{2} \sum_{j=1}^{m_i} w_j^{(i)} \boldsymbol{\delta}_j^{(i)\top} \boldsymbol{S}^{(i)}(\boldsymbol{\theta}) \boldsymbol{\delta}_j^{(i)} + \mathcal{O}(\epsilon^3).$$

Proof. Let $\overline{\boldsymbol{\theta}}^{(i)} = \{\boldsymbol{\theta}^{(i)}, \cdots, \boldsymbol{\theta}^{(i)}\}$ such that $|\overline{\boldsymbol{\theta}}^{(i)}| = m_i$. This is we split the i-th Gaussian into m_i off-springs with parameters identical to the original one, or namely we let $\epsilon = 0$. If we replace $\boldsymbol{\vartheta}^{(i)}$ with $\overline{\boldsymbol{\theta}}^{(i)}$, it holds that:

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\theta}^{(\backslash i)}, \overline{\boldsymbol{\theta}}^{(i)}, \boldsymbol{w}^{(i)}) &= \mathbb{E}_{\Pi, \boldsymbol{x} \sim \mathcal{D}(\mathcal{X})} \left[\ell \left(\sigma_\Pi(\boldsymbol{x}; \boldsymbol{\theta}^{(1)}), \cdots, \sum_{j=1}^{m_i} w_j^{(i)} \sigma_\Pi(\boldsymbol{x}; \boldsymbol{\theta}^{(i)}), \cdots, \sigma_\Pi(\boldsymbol{x}; \boldsymbol{\theta}^{(n)}) \right) \right] \\
&= \mathbb{E}_{\Pi, \boldsymbol{x} \sim \mathcal{D}(\mathcal{X})} \left[\ell \left(\sigma_\Pi(\boldsymbol{x}; \boldsymbol{\theta}^{(1)}), \cdots, \sigma_\Pi(\boldsymbol{x}; \boldsymbol{\theta}^{(i)}), \cdots, \sigma_\Pi(\boldsymbol{x}; \boldsymbol{\theta}^{(n)}) \right) \right] = \mathcal{L}(\boldsymbol{\theta}),
\end{aligned}$$

and

$$\partial_{\boldsymbol{\vartheta}_j^{(i)}} \mathcal{L}(\boldsymbol{\theta}^{(\backslash i)}, \overline{\boldsymbol{\theta}}^{(i)}, \boldsymbol{w}^{(i)}) = \partial_{\boldsymbol{\vartheta}_j^{(i)}} \left. \mathcal{L}(\boldsymbol{\theta}^{(\backslash i)}, \boldsymbol{\vartheta}^{(i)}, \boldsymbol{w}^{(i)}) \right|_{\epsilon=0}.$$

By Taylor expansion,

$$\begin{aligned}
&\mathcal{L}(\boldsymbol{\theta}^{(\backslash i)}, \boldsymbol{\vartheta}^{(i)}, \boldsymbol{w}^{(i)}) - \mathcal{L}(\boldsymbol{\theta}) = \mathcal{L}(\boldsymbol{\theta}^{(\backslash i)}, \boldsymbol{\vartheta}^{(i)}, \boldsymbol{w}^{(i)}) - \mathcal{L}(\boldsymbol{\theta}^{(\backslash i)}, \overline{\boldsymbol{\theta}}^{(i)}, \boldsymbol{w}^{(i)}) \\
&= \sum_{j=1}^{m_i} \left. \mathcal{L}(\boldsymbol{\theta}^{(\backslash i)}, \boldsymbol{\vartheta}^{(i)}, \boldsymbol{w}^{(i)}) \right|_{\epsilon=0}^\top (\boldsymbol{\vartheta}_j^{(i)} - \boldsymbol{\theta}^{(i)}) \\
&\quad + \sum_{j, j' \in [m_i]} (\boldsymbol{\vartheta}_j^{(i)} - \boldsymbol{\theta}^{(i)})^\top \left. \partial_{\boldsymbol{\vartheta}_j^{(i)} \boldsymbol{\vartheta}_{j'}^{(i)}} \mathcal{L}(\boldsymbol{\theta}^{(\backslash i)}, \boldsymbol{\vartheta}^{(i)}, \boldsymbol{w}^{(i)}) \right|_{\epsilon=0} (\boldsymbol{\vartheta}_{j'}^{(i)} - \boldsymbol{\theta}^{(i)}) + \mathcal{O}(\epsilon^3).
\end{aligned}$$

By Lemma 5 and 3:

$$\begin{aligned}
&\mathcal{L}(\boldsymbol{\theta}^{(\backslash i)}, \boldsymbol{\vartheta}^{(i)}, \boldsymbol{w}^{(i)}) - \mathcal{L}(\boldsymbol{\theta}) \\
&= \epsilon \sum_{j=1}^{m_i} w_j^{(i)} \partial_{\boldsymbol{\theta}^{(i)}} \mathcal{L}(\boldsymbol{\theta})^\top \boldsymbol{\Delta}_j^{(i)} + \frac{\epsilon^2}{2} \sum_{j=1}^{m_i} \boldsymbol{\Delta}_j^{(i)\top} \left(w_j^{(i)} \boldsymbol{S}^{(i)}(\boldsymbol{\theta}) + w_j^{(i)^2} \boldsymbol{T}^{(i)}(\boldsymbol{\theta}) \right) \boldsymbol{\Delta}_j^{(i)} \\
&\quad + \frac{\epsilon^2}{2} \sum_{j, j' \in [m_i], j \neq j'} w_j^{(i)} w_{j'}^{(i)} \boldsymbol{\Delta}_j^{(i)\top} \boldsymbol{T}^{(i)}(\boldsymbol{\theta}) \boldsymbol{\Delta}_{j'}^{(i)} + \mathcal{O}(\epsilon^3) \\
&= \epsilon \sum_{j=1}^{m_i} w_j^{(i)} \partial_{\boldsymbol{\theta}^{(i)}} \mathcal{L}(\boldsymbol{\theta})^\top \boldsymbol{\Delta}_j^{(i)} + \frac{\epsilon^2}{2} \sum_{j=1}^{m_i} w_j^{(i)} \boldsymbol{\Delta}_j^{(i)\top} \boldsymbol{S}^{(i)}(\boldsymbol{\theta}) \boldsymbol{\Delta}_j^{(i)} \\
&\quad + \frac{\epsilon^2}{2} \sum_{j, j' \in [m_i]} w_j^{(i)} w_{j'}^{(i)} \boldsymbol{\Delta}_j^{(i)\top} \boldsymbol{T}^{(i)}(\boldsymbol{\theta}) \boldsymbol{\Delta}_j^{(i)} + \mathcal{O}(\epsilon^3) \\
&= \epsilon \partial_{\boldsymbol{\theta}^{(i)}} \mathcal{L}(\boldsymbol{\theta})^\top \boldsymbol{\mu}^{(i)} + \frac{\epsilon^2}{2} \sum_{j=1}^{m_i} w_j^{(i)} \boldsymbol{\Delta}_j^{(i)\top} \boldsymbol{S}^{(i)}(\boldsymbol{\theta}) \boldsymbol{\Delta}_j^{(i)} + \boldsymbol{\mu}^{(i)^\top} \boldsymbol{T}^{(i)}(\boldsymbol{\theta}) \boldsymbol{\mu}^{(i)} + \mathcal{O}(\epsilon^3) \\
&= \epsilon \partial_{\boldsymbol{\theta}^{(i)}} \mathcal{L}(\boldsymbol{\theta})^\top \boldsymbol{\mu}^{(i)} + \frac{\epsilon^2}{2} \boldsymbol{\mu}^{(i)^\top} \left(\boldsymbol{S}^{(i)}(\boldsymbol{\theta}) + \boldsymbol{T}^{(i)}(\boldsymbol{\theta}) \right) \boldsymbol{\mu}^{(i)} \\
&\quad + \frac{\epsilon^2}{2} \sum_{j=1}^{m_i} w_j^{(i)} \left(\boldsymbol{\Delta}_j^{(i)\top} \boldsymbol{S}^{(i)}(\boldsymbol{\theta}) \boldsymbol{\Delta}_j^{(i)} - \boldsymbol{\mu}^{(i)\top} \boldsymbol{S}^{(i)}(\boldsymbol{\theta}) \boldsymbol{\mu}^{(i)} \right) + \mathcal{O}(\epsilon^3) \\
&= \epsilon \partial_{\boldsymbol{\theta}^{(i)}} \mathcal{L}(\boldsymbol{\theta})^\top \boldsymbol{\mu}^{(i)} + \frac{\epsilon^2}{2} \boldsymbol{\mu}^{(i)^\top} \partial^2_{\boldsymbol{\theta}^{(i)}\boldsymbol{\theta}^{(i)}} \mathcal{L}(\boldsymbol{\theta}) \boldsymbol{\mu}^{(i)} + \frac{\epsilon^2}{2} \sum_{j=1}^{m_i} w_j^{(i)} \left(\boldsymbol{\Delta}_j^{(i)\top} \boldsymbol{S}^{(i)}(\boldsymbol{\theta}) \boldsymbol{\Delta}_j^{(i)} - \boldsymbol{\mu}^{(i)\top} \boldsymbol{S}^{(i)}(\boldsymbol{\theta}) \boldsymbol{\mu}^{(i)} \right) + \mathcal{O}(\epsilon^3).
\end{aligned}$$

Finally, we conclude the proof by showing that:

$$\begin{aligned}
&\sum_{j=1}^{m_i} w_j^{(i)} \left(\boldsymbol{\Delta}_j^{(i)\top} \boldsymbol{S}^{(i)}(\boldsymbol{\theta}) \boldsymbol{\Delta}_j^{(i)} - \boldsymbol{\mu}^{(i)\top} \boldsymbol{S}^{(i)}(\boldsymbol{\theta}) \boldsymbol{\mu}^{(i)} \right) \\
&= \sum_{j=1}^{m_i} w_j^{(i)} \boldsymbol{\Delta}_j^{(i)\top} \boldsymbol{S}^{(i)}(\boldsymbol{\theta}) \boldsymbol{\Delta}_j^{(i)} + \boldsymbol{\mu}^{(i)\top} \boldsymbol{S}^{(i)}(\boldsymbol{\theta}) \boldsymbol{\mu}^{(i)} - 2 \left(\sum_{j=1}^{m_i} w_j^{(i)} \boldsymbol{\Delta}_j^{(i)} \right)^\top \boldsymbol{S}^{(i)}(\boldsymbol{\theta}) \boldsymbol{\mu}^{(i)} \\
&= \sum_{j=1}^{m_i} w_j^{(i)} \left((\boldsymbol{\Delta}_j^{(i)} - \boldsymbol{\mu}^{(i)})^\top \boldsymbol{S}^{(i)}(\boldsymbol{\theta}) (\boldsymbol{\Delta}_j^{(i)} - \boldsymbol{\mu}^{(i)}) \right) = \sum_{j=1}^{m_i} w_j^{(i)} {\boldsymbol{\delta}_j^{(i)}}^\top \boldsymbol{S}^{(i)}(\boldsymbol{\theta}) \boldsymbol{\delta}_j^{(i)}.
\end{aligned}$$

□

C.4. Deriving Hessian of Gaussian

In Sec. 4.4, we discussed that SteepGS requires the computation of Hessian matrices for $\sigma_\Pi(\boldsymbol{x}; \boldsymbol{\theta})$. We make the following simplifications: *(i)* We only consider position parameters as the optimization variable when computing the steepest descent directions. *(ii)* Although other variables may have a dependency on the mean parameters, *e.g.* the projection matrix and view-dependent RGB colors, we break this dependency for ease of derivation. Now suppose we have a 3D Gaussian point with parameters $\boldsymbol{\theta} = (\boldsymbol{p}, \boldsymbol{\Sigma}, o)$, where we omit RGB colors as it can be handled similarly to opacity o. Given the affine transformation $\Pi : \boldsymbol{p} \mapsto \boldsymbol{P}\boldsymbol{p} + \boldsymbol{b}$ with $\boldsymbol{P} \in \mathbb{R}^{2\times 3}$ and $\boldsymbol{b} \in \mathbb{R}^2$, then $\sigma_\Pi(\boldsymbol{x}; \boldsymbol{\theta})$ can be expressed as:

$$\sigma_\Pi(\boldsymbol{x}; \boldsymbol{\theta}) = o \exp\left(-\frac{1}{2} (\boldsymbol{x} - \boldsymbol{P}\boldsymbol{p} - \boldsymbol{b})^\top (\boldsymbol{P}\boldsymbol{\Sigma}\boldsymbol{P}^\top)^{-1} (\boldsymbol{x} - \boldsymbol{P}\boldsymbol{p} - \boldsymbol{b}) \right) = o\mathcal{N}(\boldsymbol{x}; \boldsymbol{P}\boldsymbol{p} + \boldsymbol{b}, \boldsymbol{P}\boldsymbol{\Sigma}\boldsymbol{P}^\top).$$

Its gradient can be derived as:

$$\begin{aligned}
\nabla_{\boldsymbol{p}} \sigma_\Pi(\boldsymbol{x}; \boldsymbol{\theta}) &= o\mathcal{N}(\boldsymbol{x}; \boldsymbol{P}\boldsymbol{p} + \boldsymbol{b}, \boldsymbol{P}\boldsymbol{\Sigma}\boldsymbol{P}^\top) \nabla_{\boldsymbol{p}} \left[-\frac{1}{2} (\boldsymbol{x} - \boldsymbol{P}\boldsymbol{p} - \boldsymbol{b})^\top (\boldsymbol{P}\boldsymbol{\Sigma}\boldsymbol{P}^\top)^{-1} (\boldsymbol{x} - \boldsymbol{P}\boldsymbol{p} - \boldsymbol{b}) \right] \\
&= o\mathcal{N}(\boldsymbol{x}; \boldsymbol{P}\boldsymbol{p} + \boldsymbol{b}, \boldsymbol{P}\boldsymbol{\Sigma}\boldsymbol{P}^\top) \boldsymbol{P}^\top (\boldsymbol{P}\boldsymbol{\Sigma}\boldsymbol{P}^\top)^{-1} (\boldsymbol{x} - \boldsymbol{P}\boldsymbol{p} - \boldsymbol{b}).
\end{aligned}$$

Now we can compute the Hessian matrix as:

$$\begin{aligned}
\nabla^2_{\boldsymbol{p}} \sigma_\Pi(\boldsymbol{x}; \boldsymbol{\theta}) &= \boldsymbol{P}^\top (\boldsymbol{P}\boldsymbol{\Sigma}\boldsymbol{P}^\top)^{-1} (\boldsymbol{x} - \boldsymbol{P}\boldsymbol{p} + \boldsymbol{b}) \nabla_{\boldsymbol{p}} \left[o\mathcal{N}(\boldsymbol{x}; \boldsymbol{P}\boldsymbol{p} + \boldsymbol{b}, \boldsymbol{P}\boldsymbol{\Sigma}\boldsymbol{P}^\top) \right]^\top - o\mathcal{N}(\boldsymbol{x}; \boldsymbol{P}\boldsymbol{p} + \boldsymbol{b}, \boldsymbol{P}\boldsymbol{\Sigma}\boldsymbol{P}^\top) (\boldsymbol{P}\boldsymbol{\Sigma}\boldsymbol{P}^\top)^{-1} \boldsymbol{P} \\
&= o\mathcal{N}(\boldsymbol{x}; \boldsymbol{P}\boldsymbol{p} + \boldsymbol{b}, \boldsymbol{P}\boldsymbol{\Sigma}\boldsymbol{P}^\top) \boldsymbol{P}^\top (\boldsymbol{P}\boldsymbol{\Sigma}\boldsymbol{P}^\top)^{-1} (\boldsymbol{x} - \boldsymbol{P}\boldsymbol{p} - \boldsymbol{b}) (\boldsymbol{x} - \boldsymbol{P}\boldsymbol{p} - \boldsymbol{b})^\top (\boldsymbol{P}\boldsymbol{\Sigma}\boldsymbol{P}^\top)^{-1} \boldsymbol{P} \\
&\quad - o\mathcal{N}(\boldsymbol{x}; \boldsymbol{P}\boldsymbol{p} + \boldsymbol{b}, \boldsymbol{P}\boldsymbol{\Sigma}\boldsymbol{P}^\top) \boldsymbol{P}^\top (\boldsymbol{P}\boldsymbol{\Sigma}\boldsymbol{P}^\top)^{-1} \boldsymbol{P} \\
&= \sigma_\Pi(\boldsymbol{x}; \boldsymbol{\theta}) \left(\boldsymbol{P}^\top (\boldsymbol{P}\boldsymbol{\Sigma}\boldsymbol{P}^\top)^{-1} (\boldsymbol{x} - \boldsymbol{P}\boldsymbol{p} - \boldsymbol{b}) (\boldsymbol{x} - \boldsymbol{P}\boldsymbol{p} - \boldsymbol{b})^\top (\boldsymbol{P}\boldsymbol{\Sigma}\boldsymbol{P}^\top)^{-1} \boldsymbol{P} - \boldsymbol{P}^\top (\boldsymbol{P}\boldsymbol{\Sigma}\boldsymbol{P}^\top)^{-1} \boldsymbol{P} \right)
\end{aligned}$$

as desired.