Parzen-PNN Gaussian Mixture Estimator: Deep Mathematical Proof

Fabrizio Benvenuti

February 8, 2026

1 PW error mixture 1

1.1 Observation

It was observed that for mixture 1, the MSE exhibits highly non-linear behavior as a function of window size h_1:

- The MSE approaches zero at an optimal value of h_1, forming a clear minimum.
- For h_1 below the optimum (undersmoothing), MSE increases very steeply, with rapid divergence as $h_1 \to 0$.
- For h_1 above the optimum (oversmoothing), MSE still increases but with a noticeably gentler slope.
- This creates an asymmetric U-shaped curve: sharp rise on the left (small h), shallow rise on the right (large h).
- Increasing the number of samples per Gaussian reduces MSE substantially, though this effect appears less dramatic than varying h_1.
- The NLL metric shows a shallower slope in the oversmoothed region compared to MSE.

1.2 Mathematical Foundation

1.2.1 Parzen Window Estimator Definition

Given n independent and identically distributed (i.i.d.) samples $\{x_j\}_{j=1}^n$ drawn from an unknown probability density f on $\mathbb{R}^d$, the kernel density estimator (KDE) is defined as:

$$\hat{f}_h(x) = \frac{1}{n} \sum_{j=1}^{n} K_h(x - x_j),$$

where K_h is a kernel function with bandwidth parameter $h > 0$. For an isotropic Gaussian kernel:

$$K_h(u) = \frac{1}{(2\pi h^2)^{d/2}} \exp\bigl(-\|u\|^2/(2h^2)\bigr).$$

This kernel is spherically symmetric and integrates to 1, making $\hat{f}_h$ a valid probability density.

For 2D problems ($d = 2$) with bandwidth scaling $h_n = h_1/\sqrt{n-1}$:

$$\hat{f}_{h_1}(x) = \frac{1}{n} \sum_{j=1}^{n} \frac{1}{2\pi h_n^2} \exp\bigl(-\|x - x_j\|^2/(2h_n^2)\bigr).$$

1.2.2 Mean Squared Error and Bias-Variance Decomposition

To quantify estimation accuracy, we use the pointwise mean squared error:

$$\mathrm{MSE}_h(x) = \mathbb{E}\bigl[(\hat{f}_h(x) - f(x))^2\bigr],$$

where the expectation is taken over all possible random samples of size n from f.

By the fundamental bias-variance decomposition:

$$\mathrm{MSE}_h(x) = \underbrace{\big(\mathbb{E}[\hat{f}_h(x)] - f(x)\big)^2}_{\mathrm{Bias}^2(\hat{f}_h(x))} + \underbrace{\mathbb{E}\big[(\hat{f}_h(x) - \mathbb{E}[\hat{f}_h(x)])^2\big]}_{\mathrm{Var}(\hat{f}_h(x))}.$$

This decomposition reveals two sources of error:

- **Bias**: Systematic error from the smoothing operation, measured by how far the average estimate $\mathbb{E}[\hat{f}_h(x)]$ deviates from the truth $f(x)$.
- **Variance**: Random fluctuation of $\hat{f}_h(x)$ around its mean due to finite sample size.

1.2.3 Bias Term Derivation

The expected value of the KDE is:

$$\mathbb{E}[\hat{f}_h(x)] = \mathbb{E}\Big[\frac{1}{n}\sum_{j=1}^{n} K_h(x - x_j)\Big] = \int K_h(x-u) f(u)\, du.$$

This is the convolution of the kernel with the true density. Substituting $v = x - u$ gives:

$$\mathbb{E}[\hat{f}_h(x)] = \int K_h(v) f(x-v)\, dv.$$

For smooth f and sufficiently small h, we can Taylor expand $f(x-v)$ around x:

$$f(x-v) = f(x) - v^\top \nabla f(x) + \frac{1}{2} v^\top \nabla^2 f(x)\, v + O(\|v\|^3),$$

where ∇f is the gradient and $\nabla^2 f$ is the Hessian matrix of second derivatives.

The Gaussian kernel is spherically symmetric, so all odd moments vanish:

$$\int v_i K_h(v)\, dv = 0 \quad \text{for all } i.$$

For the quadratic term, the second moment is:

$$\int v_i v_j K_h(v)\, dv = h^2 \delta_{ij},$$

where $\delta_{ij} = 1$ if $i = j$ and 0 otherwise (Kronecker delta).

Substituting into the convolution:

$$\mathbb{E}[\hat{f}_h(x)] = f(x) \underbrace{\int K_h(v)\, dv}_{=1} + \frac{1}{2}\sum_{i,j} \frac{\partial^2 f}{\partial x_i \partial x_j}(x) \underbrace{\int v_i v_j K_h(v)\, dv}_{=h^2\delta_{ij}} + O(h^4).$$

The sum over i, j with δ_{ij} picks out only diagonal terms:

$$\mathbb{E}[\hat{f}_h(x)] = f(x) + \frac{h^2}{2}\sum_{i=1}^{d} \frac{\partial^2 f}{\partial x_i^2}(x) + O(h^4) = f(x) + \frac{h^2}{2}\Delta f(x) + O(h^4),$$

where $\Delta f = \mathrm{tr}(\nabla^2 f)$ is the Laplacian, the trace of the Hessian.

The bias is therefore:

$$\mathrm{Bias}(\hat{f}_h(x)) = \mathbb{E}[\hat{f}_h(x)] - f(x) = \frac{h^2}{2}\Delta f(x) + O(h^4).$$

Interpretation: The bias is proportional to h^2 and to the local curvature $\Delta f(x)$. At density peaks (modes), $\Delta f < 0$ (concave), so the KDE underestimates the peak height. In valleys, $\Delta f > 0$ (convex), causing overestimation.

1.2.4 Integrated Squared Bias

To obtain a global error measure, we integrate the squared bias over the domain:

$$\int_{\mathbb{R}^d} \text{Bias}^2(\hat{f}_h(x))\,dx = \int_{\mathbb{R}^d} \Big(\frac{h^2}{2}\Delta f(x)\Big)^2\,dx = \frac{h^4}{4}\int_{\mathbb{R}^d} (\Delta f(x))^2\,dx.$$

Define the bias constant:

$$C_b = \frac{1}{4}\int_{\mathbb{R}^d} (\Delta f(x))^2\,dx.$$

Then the integrated squared bias is:

$$\int_{\mathbb{R}^d} \text{Bias}^2(\hat{f}_h(x))\,dx = C_b h^4.$$

Key property: The bias grows as the *fourth power* of bandwidth: $\text{Bias}^2 \propto h^4$. This is a consequence of the second-order kernel (Gaussian) and smooth density.

1.2.5 Variance Term Derivation

Since the samples are i.i.d., the variance of the KDE is:

$$\text{Var}(\hat{f}_h(x)) = \text{Var}\Big(\frac{1}{n}\sum_{j=1}^{n} K_h(x - x_j)\Big) = \frac{1}{n}\text{Var}(K_h(x - X)),$$

where $X \sim f$ is a single random sample.

By definition of variance:

$$\text{Var}(K_h(x - X)) = \mathbb{E}[K_h^2(x - X)] - (\mathbb{E}[K_h(x - X)])^2.$$

The second term is $(\mathbb{E}[K_h(x - X)])^2 = (\mathbb{E}[\hat{f}_h(x)])^2 \approx f^2(x)$, which is small for typical densities.

For the first term, we compute:

$$\mathbb{E}[K_h^2(x - X)] = \int K_h^2(x - u)f(u)\,du.$$

For the Gaussian kernel:

$$K_h(v) = \frac{1}{(2\pi h^2)^{d/2}}\exp\big(-\|v\|^2/(2h^2)\big).$$

The square is:

$$K_h^2(v) = \frac{1}{(2\pi h^2)^{d}}\exp\big(-\|v\|^2/h^2\big).$$

This can be written as:

$$K_h^2(v) = \frac{1}{(2\pi h^2)^{d/2}}\cdot\frac{1}{(2\pi(h/\sqrt{2})^2)^{d/2}}\exp\big(-\|v\|^2/(2(h/\sqrt{2})^2)\big) = \frac{1}{(2\pi h^2)^{d/2}}K_{h/\sqrt{2}}(v),$$

where $K_{h/\sqrt{2}}$ is a Gaussian kernel with bandwidth $h/\sqrt{2}$.

Therefore:

$$\mathbb{E}[K_h^2(x - X)] = \frac{1}{(2\pi h^2)^{d/2}}\int K_{h/\sqrt{2}}(x - u)f(u)\,du \approx \frac{f(x)}{(2\pi h^2)^{d/2}}.$$

The approximation holds when f varies slowly on the scale of h (which is valid for small h).

Thus:

$$\text{Var}(\hat{f}_h(x)) \approx \frac{1}{n}\cdot\frac{f(x)}{(2\pi h^2)^{d/2}}.$$

For $d = 2$:

$$\text{Var}(\hat{f}_h(x)) \approx \frac{f(x)}{2\pi n h^2}.$$

1.2.6 Integrated Variance

Integrating over $\mathbb{R}^2$:

$$\int_{\mathbb{R}^2} \mathrm{Var}(\hat{f}_h(x))\,dx \approx \frac{1}{2\pi n h^2} \int_{\mathbb{R}^2} f(x)\,dx = \frac{1}{2\pi n h^2}.$$

Define the variance constant:

$$C_v = \frac{1}{2\pi}.$$

Then the integrated variance is:

$$\int_{\mathbb{R}^2} \mathrm{Var}(\hat{f}_h(x))\,dx = \frac{C_v}{n h^2}.$$

Key property: The variance decays as h^{-2} (inverse square of bandwidth) and as n^{-1} (inverse of sample size).

1.2.7 Integrated Mean Squared Error (IMSE)

Combining the bias and variance contributions:

$$\mathrm{IMSE}(h) = \int_{\mathbb{R}^2} \mathrm{MSE}_h(x)\,dx = \underbrace{C_b h^4}_{\text{Bias}^2} + \underbrace{\frac{C_v}{n h^2}}_{\text{Variance}}.$$

This is the fundamental formula governing KDE performance. It reveals the bias-variance tradeoff:

- Large h: Bias term $C_b h^4$ dominates $\rightarrow$ oversmoothing.
- Small h: Variance term $C_v/(nh^2)$ dominates $\rightarrow$ undersmoothing.

1.3 Asymmetric U-Shape: Mathematical Explanation

1.3.1 Behavior for Small h (Undersmoothing)

When h is much smaller than the optimal value, the variance term dominates:

$$\mathrm{IMSE}(h) \approx \frac{C_v}{n h^2} \quad \text{for small } h.$$

Taking the derivative with respect to h:

$$\frac{d}{dh}\mathrm{IMSE}(h) \approx -\frac{2C_v}{n h^3} < 0.$$

The magnitude of this derivative is:

$$\left|\frac{d(\mathrm{IMSE})}{dh}\right| = \frac{2C_v}{n h^3} \propto h^{-3}.$$

Key insight: As $h \to 0$, the slope diverges as h^{-3}, creating an extremely steep rise on the left side of the U-curve. This is a *power-law divergence*, not exponential, but it is very rapid for small h.

Physical interpretation: With very small bandwidth, each kernel becomes a sharp spike. The estimate $\hat{f}_h(x)$ is dominated by random fluctuations (high variance), placing narrow peaks at sample locations and near-zero density elsewhere. This creates large squared errors averaged over the domain.

1.3.2 Behavior for Large h (Oversmoothing)

When h is much larger than the optimal value, the bias term dominates:

$$\mathrm{IMSE}(h) \approx C_b h^4 \quad \text{for large } h.$$

Taking the derivative:

$$\frac{d}{dh}\mathrm{IMSE}(h) \approx 4C_b h^3 > 0.$$

Key insight: The slope grows as h^3, which is much gentler than the h^{-3} divergence in the undersmoothed region. For moderate values of h around the optimum, h^3 grows slowly compared to h^{-3} blowing up.

Physical interpretation: With large bandwidth, each kernel spreads probability mass over a wide area, causing the estimate to be overly smooth. Peaks are flattened and valleys are filled in. The squared error increases as we deviate from the true density, but this increase is polynomial (h^4), not the inverse power-law explosion of the variance term.

1.3.3 Optimal Bandwidth

To find the bandwidth minimizing IMSE, we set the derivative to zero:

$$\frac{d}{dh}\text{IMSE}(h) = 4C_b h^3 - \frac{2C_v}{nh^3} = 0.$$

Solving for h:

$$4C_b h^3 = \frac{2C_v}{nh^3} \quad \Rightarrow \quad h^6 = \frac{C_v}{2nC_b} \quad \Rightarrow \quad h^{\text{opt}}_{\text{IMSE}} = \Big(\frac{C_v}{2nC_b}\Big)^{1/6}.$$

For $d = 2$ with $C_v = 1/(2\pi)$:

$$h^{\text{opt}}_{\text{IMSE}} = \Big(\frac{1}{4\pi n C_b}\Big)^{1/6} \propto n^{-1/6}.$$

At this optimal bandwidth, both bias and variance contribute equally to IMSE. Substituting back:

$$\text{IMSE}_{\min} = C_b (h^{\text{opt}})^4 + \frac{C_v}{n(h^{\text{opt}})^2}.$$

From the optimality condition $h^6 = C_v/(2nC_b)$, we have $(h^{\text{opt}})^4 = (C_v/(2nC_b))^{2/3}$ and $(h^{\text{opt}})^2 = (C_v/(2nC_b))^{1/3}$. After algebra:

$$\text{IMSE}_{\min} \propto n^{-2/3}.$$

1.4 Sample Size Effect: $n^{-2/3}$ Scaling

1.4.1 Why Sample Size Has Less Visual Impact Than Bandwidth

From the IMSE formula:

$$\text{IMSE}(h) = C_b h^4 + \frac{C_v}{nh^2},$$

we can analyze sensitivity to each parameter.

Sensitivity to h Taking the partial derivative:

$$\frac{\partial}{\partial h}\text{IMSE}(h) = 4C_b h^3 - \frac{2C_v}{nh^3}.$$

Near the optimum where $4C_b h^3 \approx 2C_v/(nh^3)$, the magnitude of change is dominated by the local curvature. A 10% change in h can produce a substantial change in IMSE because both terms respond strongly (one increases, the other decreases).

Sensitivity to n Taking the partial derivative with respect to n:

$$\frac{\partial}{\partial n}\text{IMSE}(h) = -\frac{C_v}{n^2 h^2} < 0.$$

The improvement from increasing n is:

$$\Delta(\text{IMSE}) \approx -\frac{C_v}{n^2 h^2}\Delta n.$$

At the optimal bandwidth $h^{\text{opt}} \propto n^{-1/6}$:

$$\text{IMSE}_{\min}(n) \propto n^{-2/3}.$$

Doubling the sample size ($n \to 2n$) reduces IMSE by a factor:

$$\frac{\text{IMSE}_{\min}(2n)}{\text{IMSE}_{\min}(n)} = (1/2)^{2/3} \approx 0.63.$$

This is a 37% reduction, which is significant but not dramatic.

In contrast, moving away from the optimal h by a factor of 2 (e.g., $h \to 2h$ in the oversmoothed region) increases IMSE by:

$$\frac{C_b(2h)^4}{C_b h^4} = 16.$$

Similarly, halving h in the undersmoothed region increases the variance term by $4\times$.

Conclusion: Changes in h produce much larger relative changes in IMSE than proportional changes in n, making bandwidth variations more visually apparent in error plots.

1.4.2 Why the Effect Is Called "Exponential-Like" Despite Being Power-Law

The observation that "samples per Gaussian seem to have an exponential impact on reducing MSE" is slightly imprecise mathematically. The true relationship is a power law:

$$\text{MSE}_{\min} \propto n^{-2/3}.$$

However, on a linear scale, this can *appear* exponential-like because:

- Power-law decay $n^{-\alpha}$ with $\alpha > 0$ is rapid for small n and slows as n increases.
- On a log-log plot, $n^{-2/3}$ appears as a straight line with slope $-2/3$.
- On a linear plot (MSE vs n), the curve is concave, resembling exponential decay qualitatively.

The key distinction: exponential decay would be $\propto \exp(-\beta n)$, which converges to zero much faster than any power law. The $n^{-2/3}$ scaling is slower but still provides substantial improvements.

1.5 Negative Log-Likelihood (NLL) vs MSE: Why Shallower Slope

1.5.1 Definition of NLL Objective

Given held-out validation samples $\{y_i\}_{i=1}^m$, the negative log-likelihood is:

$$\text{NLL}(h) = -\frac{1}{m}\sum_{i=1}^{m} \log \hat{f}_h(y_i).$$

This measures how well the estimate $\hat{f}_h$ assigns probability to observed data points. Lower NLL means higher likelihood.

1.5.2 Relationship to Kullback-Leibler Divergence

Minimizing NLL is equivalent to minimizing the expected Kullback-Leibler (KL) divergence:

$$\mathbb{E}_y[-\log \hat{f}_h(y)] = \int f(x)(-\log \hat{f}_h(x))\, dx = \text{KL}(f \| \hat{f}_h) + \text{const},$$

where the constant is the entropy of f, independent of $\hat{f}_h$.

1.5.3 Why NLL Is Less Sensitive to Oversmoothing Than MSE

Consider the behavior of the two objectives in the oversmoothed region (large h):

MSE in oversmoothed region For large h, the bias dominates:

$$\hat{f}_h(x) - f(x) \approx \frac{h^2}{2}\Delta f(x).$$

At density peaks (modes), $\Delta f < 0$, so $\hat{f}_h$ underestimates f linearly in h^2. In valleys, $\Delta f > 0$, causing overestimation. Squaring these errors:

$$\mathrm{MSE}_h(x) \approx \Big(\frac{h^2}{2}\Delta f(x)\Big)^2 \propto h^4.$$

Integrated over the domain:

$$\mathrm{IMSE}(h) \approx C_b h^4.$$

The slope is $d(\mathrm{IMSE})/dh \approx 4C_b h^3$, growing cubically with h.

NLL in oversmoothed region The log-likelihood at a point x is:

$$\log \hat{f}_h(x) = \log f(x) + \log\Big(1 + \frac{\hat{f}_h(x) - f(x)}{f(x)}\Big).$$

For small relative errors $|\hat{f}_h - f| \ll f$, Taylor expansion gives:

$$\log \hat{f}_h(x) \approx \log f(x) + \frac{\hat{f}_h(x) - f(x)}{f(x)} - \frac{1}{2}\Big(\frac{\hat{f}_h(x) - f(x)}{f(x)}\Big)^2.$$

The negative log-likelihood is:

$$-\log \hat{f}_h(x) \approx -\log f(x) - \frac{\hat{f}_h(x) - f(x)}{f(x)} + \frac{1}{2}\Big(\frac{\hat{f}_h(x) - f(x)}{f(x)}\Big)^2.$$

The integrated NLL (expected over f) is:

$$\int f(x)(-\log \hat{f}_h(x))\,dx \approx \int f(x)(-\log f(x))\,dx + \int \frac{(\hat{f}_h(x) - f(x))^2}{2f(x)}\,dx.$$

The first term is constant (entropy of f). The second term is the χ^2 divergence, which grows as:

$$\int \frac{(\hat{f}_h(x) - f(x))^2}{f(x)}\,dx \propto h^4.$$

However, the key difference is the *weighting* by $1/f(x)$:

- MSE gives equal weight to all regions of space: $\int (\hat{f}_h - f)^2\,dx$.
- NLL downweights low-density regions: $\int (\hat{f}_h - f)^2/f\,dx$.

In the oversmoothed regime, the KDE overestimates density in tails and valleys (low-density regions). These large absolute errors contribute heavily to MSE but are divided by small $f(x)$ in the NLL formula, reducing their impact.

Conversely, NLL is highly sensitive to underestimation at high-density regions (modes), where $f(x)$ is large. Oversmoothing slightly flattens peaks, but the relative error $(\hat{f}_h - f)/f$ remains moderate because both numerator and denominator are large.

1.5.4 Logarithmic Compression

The logarithm compresses multiplicative errors:

- If $\hat{f}_h(x) = 2f(x)$ ($2\times$ overestimation), $\log \hat{f}_h - \log f = \log 2 \approx 0.69$.
- If $\hat{f}_h(x) = 4f(x)$ ($4\times$ overestimation), $\log \hat{f}_h - \log f = \log 4 \approx 1.39$.

Doubling the error only increases the log-error by a constant amount, whereas MSE squares the error:

- $(\hat{f}_h - f)^2 = (f)^2$ for $2\times$ error.
- $(\hat{f}_h - f)^2 = (3f)^2 = 9f^2$ for $4\times$ error.

This logarithmic compression makes NLL less sensitive to large absolute deviations in low-density regions, which dominate MSE growth in the oversmoothed regime.

1.5.5 Quantitative Comparison

For large h, both objectives grow as h^4 asymptotically, but with different constants:

$$\text{IMSE}(h) \approx C_b h^4, \qquad \text{NLL}(h) \approx D_b h^4,$$

where D_b involves the χ^2-weighted curvature integral:

$$D_b \propto \int \frac{(\Delta f(x))^2}{f(x)}\, dx.$$

The ratio D_b/C_b depends on the distribution of curvature relative to density. For typical smooth mixtures, $D_b < C_b$ because:

- Large $|\Delta f|$ occurs at peaks (high f) and valleys (low f).
- At peaks, dividing by large f reduces the contribution.
- At valleys, both numerator and denominator are small, but the logarithmic nature of NLL compresses the impact.

Therefore, the slope $d(\text{NLL})/dh \approx 4D_b h^3$ is smaller than $d(\text{IMSE})/dh \approx 4C_b h^3$ in the oversmoothed region.

1.6 Summary: From Observation to Mathematical Proof

1.6.1 Chain of Reasoning

1. **Observed**: MSE exhibits an asymmetric U-shape with steep left side (undersmoothing), gentle right side (oversmoothing), and minimum at optimal h_1.
2. **IMSE decomposition**: For 2D KDE with Gaussian kernel,

$$\text{IMSE}(h) = C_b h^4 + \frac{C_v}{nh^2}.$$

3. **Undersmoothing slope**: For small h, variance dominates:

$$\text{IMSE}(h) \approx \frac{C_v}{nh^2}, \quad \frac{d(\text{IMSE})}{dh} \approx -\frac{2C_v}{nh^3} \propto -h^{-3}.$$

   The slope diverges as $h \to 0$, creating the steep left side.

4. **Oversmoothing slope**: For large h, bias dominates:

$$\text{IMSE}(h) \approx C_b h^4, \quad \frac{d(\text{IMSE})}{dh} \approx 4C_b h^3 \propto h^3.$$

   The slope grows polynomially, much gentler than the h^{-3} divergence.

5. **Optimum**: Setting $4C_b h^3 = 2C_v/(nh^3)$ yields

$$h^{\text{opt}} \propto n^{-1/6}, \quad \text{IMSE}_{\text{min}} \propto n^{-2/3}.$$

6. **Sample size effect**: Increasing n reduces IMSE by $n^{-2/3}$ (power-law), but local changes in h around the optimum produce larger relative changes (factor of 4 or 16 for $2\times$ bandwidth shift), making bandwidth effects more visually prominent.
7. **NLL shallower slope**: NLL uses $1/f(x)$ weighting and logarithmic compression, downweighting large absolute errors in low-density regions that dominate MSE in the oversmoothed regime. Hence $d(\text{NLL})/dh < d(\text{IMSE})/dh$ for large h.

This mathematical framework explains all observed behaviors: the asymmetric U-shape arises from the competing h^4 bias and h^{-2} variance terms, sample size provides power-law improvements dominated by local bandwidth sensitivity, and NLL's logarithmic structure compresses oversmoothing errors relative to MSE.

2 PW error mixture 2

2.1 Observation

It was noticed that the MSE slope in the oversmoothed region (large h) is less steep for mixture 2 compared to mixture 1. Specifically:

- When h exceeds the MSE-optimal bandwidth, mixture 1's MSE increases rapidly.
- Mixture 2's MSE also increases in the oversmoothed region, but the rate of growth is noticeably gentler.
- This suggests that oversmoothing causes less error when modes are closer together (mixture 2) than when they are well-separated (mixture 1).

2.2 Mathematical Foundation

2.2.1 Kernel Density Estimator

Given n i.i.d. samples $\{x_j\}_{j=1}^n$ from an unknown density f on $\mathbb{R}^d$, the Parzen window (KDE) estimate is:

$$\hat{f}_h(x) = \frac{1}{n}\sum_{j=1}^{n} K_h(x - x_j),$$

where K_h is an isotropic Gaussian kernel:

$$K_h(u) = \frac{1}{(2\pi h^2)^{d/2}} \exp\bigl(-\|u\|^2/(2h^2)\bigr).$$

The bandwidth h controls the kernel width. In this work, for 2D problems ($d = 2$):

$$\hat{f}_{h_1}(x) = \frac{1}{n}\sum_{j=1}^{n} \frac{1}{2\pi h_n^2} \exp\bigl(-\|x - x_j\|^2/(2h_n^2)\bigr), \quad h_n = \frac{h_1}{\sqrt{n-1}}.$$

2.2.2 Bias-Variance Decomposition

The mean squared error at a point x decomposes as:

$$\mathrm{MSE}_h(x) = \mathbb{E}\bigl[(\hat{f}_h(x) - f(x))^2\bigr] = \mathrm{Bias}^2(\hat{f}_h(x)) + \mathrm{Var}(\hat{f}_h(x)).$$

The expected value of the KDE is a convolution:

$$\mathbb{E}[\hat{f}_h(x)] = \int K_h(x - u) f(u)\, du.$$

For smooth f and small h, Taylor expansion of $f(x - v)$ around x gives:

$$\mathbb{E}[\hat{f}_h(x)] = f(x) + \frac{h^2}{2}\Delta f(x) + O(h^4),$$

where $\Delta f = \sum_{i=1}^{d} \frac{\partial^2 f}{\partial x_i^2}$ is the Laplacian (trace of the Hessian). This measures the local curvature of the density.

The bias is therefore:

$$\mathrm{Bias}(\hat{f}_h(x)) = \frac{h^2}{2}\Delta f(x) + O(h^4).$$

2.2.3 Integrated Mean Squared Error (IMSE)

Integrating the squared bias over the domain:

$$\int_{\mathbb{R}^d} \mathrm{Bias}^2(\hat{f}_h(x))\, dx = \frac{h^4}{4}\int_{\mathbb{R}^d} (\Delta f(x))^2\, dx \equiv C_b h^4,$$

where the bias constant is:

$$C_b = \frac{1}{4} \int_{\mathbb{R}^d} (\Delta f(x))^2 \, dx.$$

For the variance term, since samples are i.i.d.:

$$\mathrm{Var}(\hat{f}_h(x)) \approx \frac{f(x)}{n(2\pi h^2)^{d/2}}.$$

Integrating over $\mathbb{R}^d$ (for $d = 2$):

$$\int_{\mathbb{R}^2} \mathrm{Var}(\hat{f}_h(x)) \, dx \approx \frac{1}{2\pi n h^2} \equiv \frac{C_v}{nh^2},$$

where $C_v = 1/(2\pi)$.

Combining bias and variance:

$$\mathrm{IMSE}(h) = C_b h^4 + \frac{C_v}{nh^2}.$$

2.3 MSE Behavior in the Oversmoothed Region

2.3.1 Dominance of Bias Term

In the oversmoothed region (large h), the bias term $C_b h^4$ grows as the fourth power of h, while the variance term $C_v/(nh^2)$ decreases as h^{-2}. For sufficiently large h:

$$\mathrm{IMSE}(h) \approx C_b h^4.$$

Taking the derivative with respect to h:

$$\frac{d}{dh}\mathrm{IMSE}(h) \approx 4C_b h^3.$$

Key insight: The steepness (slope) of the MSE curve in the oversmoothed region is directly proportional to the bias constant C_b:

$$\text{Slope in oversmoothed region} \propto C_b.$$

A larger C_b means steeper MSE growth when oversmoothing. A smaller C_b means more gradual MSE increase.

2.3.2 Geometric Interpretation of C_b

The bias constant depends on the integrated squared Laplacian:

$$C_b = \frac{1}{4} \int_{\mathbb{R}^2} (\Delta f(x))^2 \, dx.$$

The Laplacian $\Delta f(x)$ measures curvature:

- At density peaks (modes), $\Delta f < 0$ (negative curvature, concave).
- In low-density regions (troughs), Δf can be positive (convex).
- Strong transitions from high to low density produce large $|\Delta f|$.

For a Gaussian mixture $f(x) = \sum_{k=1}^{K} \alpha_k \mathcal{N}(x; \mu_k, \Sigma_k)$:

$$\Delta f(x) = \sum_{k=1}^{K} \alpha_k \Delta \mathcal{N}(x; \mu_k, \Sigma_k).$$

2.4 Effect of Mode Separation on C_b

2.4.1 Well-Separated Modes (Mixture 1)

When Gaussian components are well-separated (far apart relative to their widths):

- Each mode creates a region of strong negative curvature at its center.
- Between modes, the density drops to near-zero, creating deep valleys (troughs).
- The transition from high-density peaks to low-density troughs involves rapid changes in curvature.
- These rapid spatial variations produce large values of $|\Delta f(x)|$ in transition regions.
- Consequence: $\int (\Delta f)^2 \, dx$ is *large*.

Physical interpretation: When oversmoothing well-separated modes, the KDE spreads probability mass from the peaks into the low-density tails and troughs between modes. Since the true density $f(x)$ is near zero in these regions, the squared error $(\hat{f}_h(x) - f(x))^2$ becomes large there. The large curvature variations $(\Delta f)^2$ directly quantify this susceptibility to oversmoothing error.

2.4.2 Close Modes (Mixture 2)

When Gaussian components are close together or overlap significantly:

- The density between modes remains elevated (no deep valleys).
- Overlapping tails "fill in" the space between peaks.
- The combined density profile is smoother with gentler transitions.
- The Laplacians of neighboring components can partially cancel when summed:

$$\Delta f = \sum_k \alpha_k \Delta \phi_k.$$

  If two Gaussians overlap, their individual Laplacians $\Delta\phi_1$ and $\Delta\phi_2$ may have opposite signs in the overlap region, reducing the magnitude of Δf.

- Consequence: $\int (\Delta f)^2 \, dx$ is *smaller*.

Physical interpretation: When oversmoothing close modes, the KDE spreads mass from one peak toward neighboring peaks. Since neighboring regions already have moderate-to-high density, the squared error $(\hat{f}_h(x) - f(x))^2$ remains relatively small. The smoother spatial curvature (smaller $|\Delta f|$) indicates lower sensitivity to oversmoothing.

2.5 Quantitative Comparison: Mixture 1 vs Mixture 2

From the mathematical relationships:

$$\begin{aligned} C_b^{(\text{mix1})} &> C_b^{(\text{mix2})} \\ &\Downarrow \\ \frac{d(\text{MSE}_{\text{mix1}})}{dh}\Big|_{h \text{ large}} &= 4C_b^{(\text{mix1})} h^3 > 4C_b^{(\text{mix2})} h^3 = \frac{d(\text{MSE}_{\text{mix2}})}{dh}\Big|_{h \text{ large}}. \end{aligned}$$

Conclusion: Mixture 1 (well-separated modes) has larger integrated squared curvature $C_b^{(\text{mix1})}$ due to deep troughs between peaks. This produces a steeper MSE slope in the oversmoothed region. Mixture 2 (closer modes) has smaller $C_b^{(\text{mix2})}$ due to smoother combined density, yielding a gentler MSE slope when oversmoothing.

2.6 Mathematical Justification: From Observation to Proof

2.6.1 Chain of reasoning

1. **Observed**: Mixture 2 shows less steep MSE growth in the oversmoothed region than mixture 1.

2. **IMSE decomposition**: For large h, $\text{IMSE}(h) \approx C_b h^4$ (bias-dominated).
3. **Slope derivation**: $\frac{d(\text{IMSE})}{dh} = 4C_b h^3 \propto C_b$ in the oversmoothed regime.
4. **Curvature definition**: $C_b = \frac{1}{4}\int (\Delta f)^2\, dx$ quantifies integrated squared Laplacian.
5. **Geometric analysis**:
   - Well-separated modes $\rightarrow$ deep troughs $\rightarrow$ large $|\Delta f| \rightarrow$ large C_b.
   - Close modes $\rightarrow$ filled valleys $\rightarrow$ small $|\Delta f| \rightarrow$ small C_b.
6. **Conclusion**: Since mixture 2 has closer modes than mixture 1, we have $C_b^{(\text{mix2})} < C_b^{(\text{mix1})}$, which directly implies less steep oversmoothed MSE slope for mixture 2.

This explains both *what* was observed (gentler MSE slope) and *why* it occurs (smaller curvature integral from closer modes).

3 PW error mixture 3

3.1 Observation

For mixture 3 (5 Gaussians), the following was observed:

- Increasing the number of sampled points has less impact on reducing MSE compared to mixtures 1 and 2.
- The NLL curve shape and optimal h_1 vary less across mixtures than the MSE curve does.
- This causes the selected mixtures to be undersmoothed (NLL-optimal h smaller than MSE-optimal h), the more Gaussians there are in the mixture.
- NLL-optimal bandwidths are consistently smaller than MSE-optimal ones, with the gap increasing for more Gaussians:

$$h^{\text{opt}}_{1,\text{MSE}}(\text{mix}_1, \text{mix}_2, \text{mix}_3) = (5.13,\ 5.54,\ 7.48), \qquad h^{\text{opt}}_{1,\text{NLL}}(\text{mix}_1, \text{mix}_2, \text{mix}_3) = (2.11,\ 3.19,\ 3.75).$$

3.2 Mathematical Foundation

3.2.1 Kernel Density Estimator Definition

Given n independent and identically distributed (i.i.d.) samples $\{x_j\}_{j=1}^n$ drawn from an unknown probability density f on $\mathbb{R}^d$, we want to estimate $f(x)$ at any point x.

The Parzen window (kernel density estimator, KDE) places a "bump" (kernel) at each sample point and averages them:

$$\hat{f}_h(x) = \frac{1}{n}\sum_{j=1}^{n} K_h(x - x_j),$$

where K_h is the kernel function. Intuitively, $\hat{f}_h(x)$ is high where many samples are nearby, and low where samples are sparse.

We use an isotropic (same width in all directions) Gaussian kernel:

$$K_h(u) = \frac{1}{(2\pi h^2)^{d/2}} \exp\bigl(-\|u\|^2/(2h^2)\bigr),$$

where $h > 0$ is the bandwidth controlling the kernel width. The normalization constant $(2\pi h^2)^{-d/2}$ ensures $\int K_h(u)\, du = 1$ (probability distribution).

For a 2D problem ($d = 2$), this becomes:

$$K_h(u) = \frac{1}{2\pi h^2} \exp\bigl(-\|u\|^2/(2h^2)\bigr).$$

In this work, the effective bandwidth shrinks with sample size: $h_n = h_1/\sqrt{n-1}$, where h_1 is the base bandwidth parameter. This gives:

$$\hat{f}_{h_1}(x) = \frac{1}{n}\sum_{j=1}^{n} \frac{1}{2\pi h_n^2} \exp\big(-\|x - x_j\|^2/(2h_n^2)\big), \quad h_n = \frac{h_1}{\sqrt{n-1}}.$$

Note: As n increases, h_n decreases, making each kernel narrower (more peaked). This means the estimate becomes sharper with more data.

3.2.2 Mean Squared Error (MSE) and IMSE

To measure how well $\hat{f}_h(x)$ approximates $f(x)$, we use the mean squared error (MSE) at a point x:

$$\mathrm{MSE}_h(x) = \mathbb{E}\big[(\hat{f}_h(x) - f(x))^2\big].$$

The expectation $\mathbb{E}[\cdot]$ is taken over all possible sets of n samples we might draw. This tells us the average squared error at x if we repeated the experiment many times.

By the bias-variance decomposition (a fundamental identity in statistics), we can write:

$$\mathrm{MSE}_h(x) = \underbrace{\big(\mathbb{E}[\hat{f}_h(x)] - f(x)\big)^2}_{\text{Bias}^2} + \underbrace{\mathbb{E}\big[(\hat{f}_h(x) - \mathbb{E}[\hat{f}_h(x)])^2\big]}_{\text{Variance}}.$$

- **Bias**: How far off the *average* estimate is from the truth. Large h oversmooths, creating bias.

The expected value of our estimate is:

$$\mathbb{E}[\hat{f}_h(x)] = \mathbb{E}\Big[\frac{1}{n}\sum_{j=1}^{n} K_h(x - x_j)\Big] = \frac{1}{n}\sum_{j=1}^{n} \mathbb{E}[K_h(x - x_j)].$$

Since each x_j is drawn from f, we have $\mathbb{E}[K_h(x - x_j)] = \int K_h(x - u) f(u)\, du$. By change of variable $v = x - u$:

$$\mathbb{E}[\hat{f}_h(x)] = \int K_h(v) f(x - v)\, dv.$$

This is the *convolution* of the kernel with the true density. If h is small and f is smooth, we can Taylor expand $f(x - v)$ around x:

$$f(x - v) = f(x) - v^\top \nabla f(x) + \frac{1}{2} v^\top \nabla^2 f(x)\, v + O(\|v\|^3).$$

Substituting and using symmetry of the Gaussian kernel (odd moments vanish: $\int v_i K_h(v)\, dv = 0$):

$$\mathbb{E}[\hat{f}_h(x)] = f(x) \underbrace{\int K_h(v)\, dv}_{=1} + \frac{1}{2}\sum_{i,j} \frac{\partial^2 f}{\partial x_i \partial x_j}(x) \underbrace{\int v_i v_j K_h(v)\, dv}_{=h^2\delta_{ij}} + O(h^4).$$

For the isotropic Gaussian kernel, $\int v_i v_j K_h(v)\, dv = h^2 \delta_{ij}$ (where $\delta_{ij} = 1$ if $i = j$, else 0). Thus:

$$\mathbb{E}[\hat{f}_h(x)] = f(x) + \frac{h^2}{2}\sum_{i=1}^{d} \frac{\partial^2 f}{\partial x_i^2}(x) + O(h^4) = f(x) + \frac{h^2}{2}\Delta f(x) + O(h^4),$$

where $\Delta f = \sum_{i=1}^{d} \frac{\partial^2 f}{\partial x_i^2}$ is the *Laplacian* (trace of the Hessian matrix), measuring local curvature.

The bias is therefore:

$$\mathrm{Bias}(\hat{f}_h(x)) = \mathbb{E}[\hat{f}_h(x)] - f(x) = \frac{h^2}{2}\Delta f(x) + O(h^4).$$

Now we square and integrate. Ignoring higher-order terms (valid for small h):

$$\int_{\mathbb{R}^2} \mathrm{Bias}^2(\hat{f}_h(x))\, dx = \int_{\mathbb{R}^2} \Big(\frac{h^2}{2}\Delta f(x)\Big)^2 dx = \frac{h^4}{4}\int_{\mathbb{R}^2} (\Delta f(x))^2\, dx.$$

Define the bias constant:

$$C_b = \frac{1}{4}\int_{\mathbb{R}^2} (\Delta f(x))^2\, dx.$$

Then the integrated squared bias is $C_b h^4$.

3.2.3 Variance Term

The variance of $\hat{f}_h(x)$ measures how much the estimate fluctuates. Since $\hat{f}_h(x) = \frac{1}{n}\sum_{j=1}^{n} K_h(x - x_j)$ and the samples are independent:

$$\mathrm{Var}(\hat{f}_h(x)) = \mathrm{Var}\Big(\frac{1}{n}\sum_{j=1}^{n} K_h(x - x_j)\Big) = \frac{1}{n^2}\sum_{j=1}^{n} \mathrm{Var}(K_h(x - x_j)) = \frac{1}{n}\mathrm{Var}(K_h(x - X)),$$

where X is a single sample from f. By definition of variance:

$$\mathrm{Var}(K_h(x - X)) = \mathbb{E}[K_h^2(x - X)] - (\mathbb{E}[K_h(x - X)])^2.$$

The second term $(\mathbb{E}[K_h(x - X)])^2 \approx f^2(x)$ is small (order $O(1)$) and dominated by the first term for small h. Let's compute $\mathbb{E}[K_h^2(x - X)]$:

$$\mathbb{E}[K_h^2(x - X)] = \int K_h^2(x - u) f(u)\, du.$$

For the Gaussian kernel:

$$K_h(v) = \frac{1}{(2\pi h^2)^{d/2}} \exp\big(-\|v\|^2/(2h^2)\big).$$

The square is:

$$K_h^2(v) = \frac{1}{(2\pi h^2)^{d}} \exp\big(-\|v\|^2/h^2\big) = \frac{1}{(2\pi h^2)^{d}} \exp\big(-\|v\|^2/(2(h/\sqrt{2})^2)\big).$$

This is proportional to a Gaussian with bandwidth $h/\sqrt{2}$. Normalizing:

$$K_h^2(v) = \frac{1}{(2\pi h^2)^{d/2}} \cdot \frac{1}{(2\pi(h/\sqrt{2})^2)^{d/2}} \exp\big(-\|v\|^2/(2(h/\sqrt{2})^2)\big) = \frac{1}{(2\pi h^2)^{d/2}} K_{h/\sqrt{2}}(v).$$

Thus:

$$\mathbb{E}[K_h^2(x - X)] = \frac{1}{(2\pi h^2)^{d/2}} \int K_{h/\sqrt{2}}(x - u) f(u)\, du \approx \frac{f(x)}{(2\pi h^2)^{d/2}}$$

(assuming f is slowly varying on scale h). Therefore:

$$\mathrm{Var}(\hat{f}_h(x)) \approx \frac{1}{n} \cdot \frac{f(x)}{(2\pi h^2)^{d/2}}.$$

For $d = 2$:

$$\mathrm{Var}(\hat{f}_h(x)) \approx \frac{f(x)}{2\pi n h^2}.$$

Integrating over $\mathbb{R}^2$ results:

$$\mathrm{IMSE}(h) = \underbrace{C_b h^4}_{\text{bias}^2} + \underbrace{\frac{C_v}{nh^2}}_{\text{variance}}.$$

This reveals the *bias-variance tradeoff*:

- Large h: bias term $C_b h^4$ dominates $\rightarrow$ oversmoothing.
- Small h: variance term $C_v/(nh^2)$ dominates $\rightarrow$ undersmoothing (noisy).

To find the optimal bandwidth, we minimize IMSE by taking the derivative with respect to h:

$$\frac{d}{dh}\mathrm{IMSE}(h) = \frac{d}{dh}\Big(C_b h^4 + C_v n^{-1} h^{-2}\Big) = 4C_b h^3 - 2C_v n^{-1} h^{-3}.$$

Setting this to zero:

$$4C_b h^3 = 2C_v n^{-1} h^{-3} \quad \Rightarrow \quad 4C_b h^6 = \frac{2C_v}{n} \quad \Rightarrow \quad h^6 = \frac{C_v}{2nC_b}.$$

Taking the sixth root:

$$h_{\mathrm{IMSE}}^{\mathrm{opt}} = \Big(\frac{C_v}{2nC_b}\Big)^{1/6}.$$

For $d = 2$, $C_v = 1/(2\pi)$, so:

$$h^{\text{opt}}_{\text{IMSE}} = \Big(\frac{1}{4\pi n C_b}\Big)^{1/6} \propto n^{-1/6} C_b^{-1/6}.$$

Important observations:

1. $h^{\text{opt}} \propto n^{-1/6}$: Optimal bandwidth shrinks slowly with sample size.
2. $h^{\text{opt}} \propto C_b^{-1/6}$: Larger curvature (larger C_b) requires *smaller* h to control bias. Smoother densities (smaller C_b) allow larger h.

Plugging h^{opt} back into IMSE:

$$\text{IMSE}_{\min} = C_b (h^{\text{opt}})^4 + \frac{C_v}{n (h^{\text{opt}})^2}.$$

Since $C_b (h^{\text{opt}})^6 = C_v/(2n)$ (from the optimality condition), we have $(h^{\text{opt}})^4 = (C_v/(2nC_b))^{2/3}$ and $(h^{\text{opt}})^2 = (C_v/(2nC_b))^{1/3}$. Substituting:

$$\text{IMSE}_{\min} = C_b \Big(\frac{C_v}{2nC_b}\Big)^{2/3} + C_v n^{-1} \Big(\frac{2nC_b}{C_v}\Big)^{1/3} \propto n^{-2/3}.$$

Thus the minimum achievable IMSE decreases as $n^{-2/3}$, not exponentially but as a power law.

3.2.4 Variance Term

Since the samples are i.i.d.,

$$\text{Var}(\hat{f}_h(x)) = \frac{1}{n}\text{Var}(K_h(x - X)) = \frac{1}{n}\Big(\mathbb{E}[K_h^2(x - X)] - (\mathbb{E}[K_h(x - X)])^2\Big).$$

For small h, $\mathbb{E}[K_h^2(x - X)] \approx K_{h/\sqrt{2}}(0) f(x) = \frac{f(x)}{(4\pi h^2)^{d/2}}$ (product of two Gaussians), so

$$\text{Var}(\hat{f}_h(x)) \approx \frac{f(x)}{n(4\pi h^2)^{d/2}}.$$

Integrating over $\mathbb{R}^2$ (with $d = 2$),

$$\int_{\mathbb{R}^2} \text{Var}(\hat{f}_h(x))\, dx \approx \frac{1}{n(4\pi h^2)} \int_{\mathbb{R}^2} f(x)\, dx = \frac{1}{4\pi n h^2} \equiv \frac{C_v}{nh^2},$$

where the variance constant is $C_v = 1/(4\pi)$.

3.2.5 IMSE Decomposition and Optimal Bandwidth

Combining the bias and variance terms,

$$\text{IMSE}(h) = C_b h^4 + \frac{C_v}{nh^2}.$$

To minimize, take the derivative with respect to h and set to zero:

$$\frac{d}{dh}\text{IMSE}(h) = 4C_b h^3 - \frac{2C_v}{nh^3} = 0 \quad \Rightarrow \quad h^6 = \frac{C_v}{2nC_b}.$$

Hence the IMSE-optimal bandwidth is

$$h^{\text{opt}}_{\text{IMSE}} = \Big(\frac{C_v}{2nC_b}\Big)^{1/6} = \Big(\frac{1}{8\pi n C_b}\Big)^{1/6} \propto n^{-1/6}.$$

The minimum IMSE scales as

$$\text{IMSE}_{\min} \propto n^{-2/3}.$$

3.3 Why Mixture 3 Has Larger $h^{\text{opt}}_{\text{MSE}}$ but Less Steep Oversmoothed Region

3.3.1 MSE Slope in the Oversmoothed Region

In the oversmoothed region (large h), the bias term dominates the variance term in the IMSE:

$$\text{IMSE}(h) \approx C_b h^4 \quad \text{for large } h.$$

Taking the derivative with respect to h:

$$\frac{d}{dh}\text{IMSE}(h) \approx 4C_b h^3.$$

Thus the *steepness* (slope) of the MSE curve in the oversmoothed region is directly proportional to C_b:

$$\text{Slope in oversmoothed region} \propto C_b.$$

Key insight: A larger bias constant C_b means a steeper increase in MSE when oversmoothing. A smaller C_b means the MSE increases more gradually with h.

3.3.2 Effect of Gaussian Configuration on Curvature

The bias constant depends on the integrated squared Laplacian:

$$C_b = \frac{1}{4}\int_{\mathbb{R}^2} (\Delta f(x))^2 \, dx.$$

For a Gaussian mixture $f(x) = \sum_{k=1}^{K} \alpha_k \mathcal{N}(x; \mu_k, \Sigma_k)$, the Laplacian is:

$$\Delta f(x) = \sum_{k=1}^{K} \alpha_k \Delta\mathcal{N}(x; \mu_k, \Sigma_k).$$

For a single 2D isotropic Gaussian $\phi(x) = \mathcal{N}(x; \mu, \sigma^2 I)$:

$$\Delta\phi(x) = \phi(x)\Big(\frac{\|x-\mu\|^2}{\sigma^4} - \frac{2}{\sigma^2}\Big).$$

At the mode center ($x = \mu$), $\Delta\phi(\mu) = -\phi(\mu) \cdot 2/\sigma^2 < 0$ (negative curvature). Away from the mode, where $\|x-\mu\|^2 > 2\sigma^2$, the Laplacian becomes positive (positive curvature).

Crucially, from the optimality condition $h^{\text{opt}}_{\text{IMSE}} = (C_v/(2nC_b))^{1/6}$, we see:

$$\text{Larger } C_b \;\Rightarrow\; \text{Smaller } h^{\text{opt}}, \qquad \text{Smaller } C_b \;\Rightarrow\; \text{Larger } h^{\text{opt}}.$$

Two competing effects on C_b:

(1) Well-separated modes increase C_b When Gaussians are *well-separated* (far apart relative to their widths):

- Each mode creates a region of strong negative curvature at its center.
- Between modes, the density drops to near-zero, creating deep "valleys" (troughs).
- The transition from high-density peaks to low-density troughs involves strong positive curvature regions (where $\Delta f > 0$).
- These rapid transitions produce large $|\Delta f|$, increasing $\int (\Delta f)^2 \, dx$.

Result: Well-separated modes $\Rightarrow$ large $C_b \Rightarrow$ *small* $h^{\text{opt}} \Rightarrow$ steep MSE slope when oversmoothing.

(2) Overlapping modes decrease C_b When Gaussians *overlap significantly* (close together or with broad widths):

- The density between modes remains elevated (no deep valleys).
- The combined density is smoother, with gentler transitions.
- Overlapping tails "fill in" the troughs, reducing the magnitude of $|\Delta f|$ in transition regions.
- The Laplacians of individual components partially cancel when summed: $\Delta f = \sum_k \alpha_k \Delta\phi_k$.

Result: Overlapping modes $\Rightarrow$ smaller $C_b \Rightarrow$ *larger* $h^{\text{opt}} \Rightarrow$ less steep MSE slope when oversmoothing.

3.3.3 Reconciling Mixture 3's Behavior

The observation shows that mixture 3 (5 Gaussians) has:

1. Larger $h^{\text{opt}}_{\text{MSE}} = 7.48$ compared to mixtures 1 and 2 (5.13 and 5.54).
2. *Less steep* MSE increase in the oversmoothed region.

Both observations are **consistent** and point to the same underlying cause: mixture 3 has *smaller* C_b than mixtures 1 and 2.

Why both behaviors indicate smaller C_b for mixture 3 From the mathematical relationships derived above:

- **Optimality**: $h^{\text{opt}}_{\text{IMSE}} = (C_v/(2nC_b))^{1/6} \propto C_b^{-1/6}$

$$\text{Smaller } C_b \;\Rightarrow\; \text{Larger } h^{\text{opt}}.$$

- **Oversmoothed slope**: $d(\text{IMSE})/dh \approx 4C_b h^3$ for large h

$$\text{Smaller } C_b \;\Rightarrow\; \text{Less steep slope.}$$

Both observed features (larger h^{opt} and gentler slope) directly follow from smaller C_b.

Physical interpretation: Overlapping Gaussians Mixture 3's smaller C_b indicates that its Gaussians *overlap more significantly* than those in mixtures 1 and 2:

Scenario: Overlapping components in mixture 3

- The 5 Gaussians in mixture 3 are positioned closer together (in relative terms) or have larger variances that cause significant overlap.
- The combined density $f(x) = \sum_{k=1}^{5} \alpha_k \phi_k(x)$ is *smoother* overall:
  - Overlapping tails fill in valleys between modes.
  - Transitions from high to low density are gentler.
  - The Laplacians $\Delta\phi_k$ of neighboring components have opposite signs in overlap regions and partially cancel in the sum $\Delta f = \sum_k \alpha_k \Delta\phi_k$.
- This reduces the integrated squared curvature: $C_b = \int (\Delta f)^2\, dx$ is *smaller* for mixture 3 despite having more components.

Quantitative consequences:

1. **Larger** $h^{\text{opt}}_{\text{IMSE}}$: From $h^{\text{opt}}_{\text{IMSE}} = (C_v/(2nC_b))^{1/6}$, a smaller C_b yields a larger optimal bandwidth:

$$C_b^{(\text{mix3})} < C_b^{(\text{mix2})} < C_b^{(\text{mix1})} \quad \Rightarrow \quad h^{\text{opt}}_{1,\text{MSE}}(\text{mix}_3) = 7.48 > h^{\text{opt}}_{1,\text{MSE}}(\text{mix}_2) = 5.54 > h^{\text{opt}}_{1,\text{MSE}}(\text{mix}_1) = 5.13.$$

   The smoother (more overlapping) structure of mixture 3 has less extreme curvature variations, requiring larger h to balance bias and variance.

2. **Less steep oversmoothed slope**: From $d(\text{IMSE})/dh \approx 4C_b h^3$ in the oversmoothed region:

$$C_b^{(\text{mix3})} < C_b^{(\text{mix1})} \quad \Rightarrow \quad \frac{d(\text{MSE}_{\text{mix3}})}{dh} < \frac{d(\text{MSE}_{\text{mix1}})}{dh} \quad \text{for large } h.$$

Mixture 3's MSE increases more gradually with oversmoothing because the density is already smooth.

3.3.4 Quantitative Intuition for Overlapping Gaussians

Consider two extreme cases in 1D for simplicity:

Case A: Well-separated Gaussians Two Gaussians with means $\mu_1 = -3\sigma$, $\mu_2 = +3\sigma$, both with variance σ^2:

- The density between them drops to nearly zero: $f(0) \approx 0$.
- The second derivative oscillates strongly: negative near peaks, large positive in the valley.
- Result: large $\int (\Delta f)^2 \, dx$.

Case B: Overlapping Gaussians Two Gaussians with means $\mu_1 = -0.5\sigma$, $\mu_2 = +0.5\sigma$, both with variance σ^2:

- The density remains high everywhere: $f(0) = \frac{1}{2}\phi_1(0) + \frac{1}{2}\phi_2(0) > 0.3 \cdot \max(f)$.
- The combined profile is nearly a single broad peak (almost Gaussian).
- The Laplacians $\Delta\phi_1$ and $\Delta\phi_2$ have opposite signs in the overlap region and partially cancel.
- Result: small $\int (\Delta f)^2 \, dx$.

For mixture 3 with 5 Gaussians: if the components have significant overlap (e.g., means within 1-2 standard deviations of each other), the combined density is much smoother than 5 isolated peaks, yielding a smaller C_b than one might naively expect from "more components = more curvature".

3.3.5 Note on Sample Allocation

With more Gaussian components, the total sample budget N is distributed among more modes. However, this does *not* directly affect the bias constant C_b, which depends only on the true density $f(x)$, not on the sampling process.

The sample size n affects the *variance* term $C_v/(nh^2)$ and determines the rate of convergence, but the *structure* of the optimal bandwidth (its dependence on C_b) is determined by the geometry of the true density.

Conclusion: The primary explanation for mixture 3's behavior is **geometric**: the 5 Gaussians overlap more significantly, creating a smoother combined density with smaller integrated squared curvature (C_b). This leads to both larger h^{opt} and gentler MSE growth when oversmoothing.

3.3.6 Why Increasing Samples Has Less Impact for Mixture 3

The minimum achievable IMSE at the optimal bandwidth scales as:

$$\text{IMSE}_{\text{min}} = C_b(h^{\text{opt}})^4 + \frac{C_v}{n(h^{\text{opt}})^2} \propto (C_b^2 C_v)^{1/3} n^{-2/3}.$$

All mixtures exhibit the same $n^{-2/3}$ convergence rate. However, mixture 3 appears to show "diminishing returns" from increased sample size for several reasons:

(1) Lower baseline error Mixture 3's smaller C_b (from overlapping Gaussians) yields a smaller proportionality constant in $(C_b^2 C_v)^{1/3}$:

- The absolute MSE for mixture 3 is lower at any given sample size compared to mixtures 1 and 2.
- The same relative reduction $(n_2/n_1)^{-2/3}$ translates to a smaller absolute MSE decrease when starting from a lower baseline.
- Visually, the MSE curve appears "flatter" even though the relative rate of improvement is identical.

(2) Sample allocation across modes With 5 Gaussians sharing the total sample budget:

- Each individual mode receives approximately $N/5$ samples.
- Per-mode estimation resolution improves more slowly compared to mixtures with fewer components receiving more samples per mode.
- While this doesn't change the theoretical $n^{-2/3}$ rate, it affects the constants and can make improvements less visually apparent.

(3) Bandwidth scaling mismatch The $h_n = h_1/\sqrt{n-1} \propto n^{-1/2}$ scaling shrinks much faster than the IMSE-optimal $n^{-1/6}$:

- As n increases, kernels become excessively sharp, increasing variance.
- This variance growth partially offsets the theoretical MSE reduction from having more samples.
- The effect is most pronounced when using fixed h_1 values that may not be optimal for the range of n considered.

Conclusion: Mixture 3 follows the same $n^{-2/3}$ convergence law but appears to show smaller improvements because (a) it starts with lower error due to smoother structure, making absolute gains smaller, and (b) the rapid $n^{-1/2}$ bandwidth shrinkage may introduce additional variance that obscures the theoretical sampling benefit.

3.4 Negative Log-Likelihood (NLL) Objective

3.4.1 Definition

Given held-out validation samples $\{y_i\}_{i=1}^m$, the average negative log-likelihood is

$$\mathrm{NLL}(h) = -\frac{1}{m}\sum_{i=1}^{m} \log \hat{f}_h(y_i).$$

Minimizing NLL is equivalent to minimizing the Kullback-Leibler divergence

$$\mathrm{KL}(f\|\hat{f}_h) = \int f(x)\log\frac{f(x)}{\hat{f}_h(x)}\,dx = -\int f(x)\log\hat{f}_h(x)\,dx + \text{const}.$$

3.4.2 Why NLL Prefers Smaller h

The log loss $-\log\hat{f}_h(y)$ is highly sensitive to underestimation at sample locations:

- If $\hat{f}_h(y) \ll f(y)$, then $-\log\hat{f}_h(y) \to +\infty$ rapidly.
- Reducing h sharpens the kernel peaks, increasing $\hat{f}_h(y)$ at data points (modes), thereby reducing NLL.
- This is true even if sharpening increases pointwise variance and L2 error in low-density regions (which MSE penalizes heavily).

Hence NLL favors *sharper* estimates (smaller h) that capture high peaks at sample locations, whereas MSE balances global L2 error and prefers smoother estimates (larger h).

3.4.3 Insensitivity of NLL to Mixture Complexity

Because NLL evaluates density only at sample locations (which concentrate in high-density modes), it is less affected by the global structure (number of modes, curvature everywhere) than MSE. Therefore:

- $h_{\mathrm{NLL}}^{\mathrm{opt}}$ changes less across mixtures: $(2.11, 3.19, 3.75)$.
- $h_{\mathrm{MSE}}^{\mathrm{opt}}$ reflects global curvature and increases more: $(5.13, 5.54, 7.48)$.

3.5 Effect of $h_n = h_1/\sqrt{n-1}$ Scaling

3.5.1 Comparison to IMSE-Optimal Scaling

The IMSE-optimal bandwidth scales as $h \propto n^{-1/6}$. In this work, the scaling is

$$h_n = \frac{h_1}{\sqrt{n-1}} \approx \frac{h_1}{\sqrt{n}} \propto n^{-1/2}.$$

This is *much faster* shrinkage than $n^{-1/6}$, causing:

- Rapid increase in kernel sharpness as n grows.
- Higher variance (more peaky estimates) for any fixed h_1.
- Amplified sensitivity to the choice of h_1.

3.5.2 Impact on NLL vs MSE

With $h_n \propto n^{-1/2}$:

- The estimate becomes very peaked, increasing pointwise variance.
- NLL benefits from sharp peaks at sample locations, so NLL-optimal h_1 stays small to exploit this.
- MSE suffers from high variance and prefers larger h_1 to smooth out the estimate, especially for complex mixtures where bias is already large.

This exacerbates the gap $h^{\text{opt}}_{1,\text{MSE}} \gg h^{\text{opt}}_{1,\text{NLL}}$, particularly for mixture 3.

3.6 Conclusion

The observed behavior for mixture 3—weaker MSE improvement with more samples, larger $h^{\text{opt}}_{\text{MSE}}$, smaller and stable $h^{\text{opt}}_{\text{NLL}}$, and increasing undersmoothing—follows from:

1. **Geometric structure**: Mixture 3's 5 Gaussians overlap more significantly than the components in mixtures 1 and 2, creating a *smoother* combined density with smaller integrated squared curvature:

   $$C_b = \frac{1}{4}\int (\Delta f)^2\, dx.$$

   Overlapping tails fill in valleys, and the Laplacians of neighboring components partially cancel, reducing C_b.

2. **Optimal bandwidth**: From $h^{\text{opt}}_{\text{IMSE}} \propto C_b^{-1/6}$, smaller C_b yields larger h^{opt}:

   $$C_b^{(\text{mix3})} < C_b^{(\text{mix1,2})} \quad \Rightarrow \quad h^{\text{opt}}_{1,\text{MSE}}(\text{mix}_3) = 7.48 > 5.54 \approx 5.13.$$

3. **Oversmoothing slope**: The MSE gradient in the oversmoothed region is $d(\text{IMSE})/dh \approx 4C_b h^3$, so smaller C_b produces a gentler slope.

4. **NLL insensitivity**: NLL evaluates $-\log \hat{f}_h$ only at sample locations (high-density regions), making it insensitive to global geometric structure. NLL favors sharp peaks (small h) regardless of mixture complexity:

   $$h^{\text{opt}}_{1,\text{NLL}}(\text{mix}_1, \text{mix}_2, \text{mix}_3) = (2.11,\ 3.19,\ 3.75) \quad \text{(relatively stable)}.$$

5. **Bandwidth scaling**: The $h_n = h_1/\sqrt{n-1} \propto n^{-1/2}$ scaling shrinks much faster than IMSE-optimal $n^{-1/6}$, amplifying variance and making NLL favor even smaller h_1 to exploit sharp peaks.

6. **Increasing undersmoothing gap**: For mixture 3, MSE requires larger h (due to small C_b) while NLL still prefers small h (sharp peaks). The gap grows:

   $$h^{\text{opt}}_{1,\text{MSE}} - h^{\text{opt}}_{1,\text{NLL}} = (3.02,\ 2.35,\ 3.73) \quad \text{for mixtures 1, 2, 3.}$$

   Using NLL-selected h produces increasingly undersmoothed estimates (high variance, insufficient smoothing) as the number of overlapping Gaussians increases.

7. **Sample size effects**: The minimum IMSE $\propto (C_b^2 C_v)^{1/3} n^{-2/3}$ decreases with n at the same rate for all mixtures. However, mixture 3's smaller baseline error (from smaller C_b) makes absolute improvements appear smaller, and the rapid $n^{-1/2}$ bandwidth shrinkage may partially offset sampling benefits.

Summary: Mixture 3's overlapping Gaussians create a smoother density (small C_b) requiring larger MSE-optimal bandwidth, while NLL remains fixated on sharp peaks at sample locations, producing increasingly suboptimal (undersmoothed) density estimates as mixture complexity increases.