Questions?

A. Asymptotics, Bias–Variance, and Rates

1. Why does squared bias scale like h^4 in KDE/local constant regression?

Because a symmetric kernel has $\int uK(u)\,du = 0$, the linear term in the Taylor expansion vanishes, leaving:

$$\text{Bias}(x) \approx \tfrac{1}{2}h^2 f''(x)\,\mu_2(K).$$

Squaring gives $O(h^4)$.

---

2. Why does $O(h^4)O(1/(nh)) = o(1/(nh))$?

Under the standard conditions $h \to 0$ and $nh \to \infty$,
$h^4 \ll 1/(nh)$.
With $h \asymp n^{-1/5}$, $h^4 \asymp n^{-4/5}$ and $1/(nh) \asymp n^{-4/5}$. Any undersmoothing drives h^4 to smaller order.

---

3. Difference between $O_p(\cdot)$ and $o_p(\cdot)$?

- $X_n = O_p(a_n)$: X_n/a_n is bounded in probability.
- $X_n = o_p(a_n)$: $X_n/a_n \to 0$ in probability.

E.g., KDE variance is $O_p((nh)^{-1/2})$; under undersmoothing, bias is $o_p(\text{SE})$.

---

4. Why are higher-order kernels rarely used?

They have negative lobes → oscillations, high variance, boundary instability, and delicate bandwidth choice. Bandwidth matters more than kernel.

---

5. Explain the curse of dimensionality geometrically.

Volume of a d-dimensional ball is proportional to h^d; effective sample size is nh^d. To keep variance small, h must be large → high bias. Optimal rates $n^{-4/(4+d)}$ degrade quickly as d grows.

---

6. Why does the optimal bandwidth in 1D scale as $n^{-1/5}$?

MSE $\approx C_1 h^4 + C_2/(nh)$.
Set derivative to zero: $h^5 \asymp 1/n \to h \asymp n^{-1/5} \to$ MSE $n^{-4/5}$.

---

7. Why is the parametric MSE rate n^{-1} unattainable?

Because nonparametric estimators estimate an infinite-dimensional function; local smoothing creates bias. The h^4–$1/(nh)$ balance forces slower $n^{-4/5}$.

---

8. Why doesn't interpolation work in nonparametric regression?

Interpolation fits noise $\to$ infinite variance. Smoothers trade small bias for large variance reduction.

---

9. Why do smoothness assumptions matter?

Without bounded derivatives, the local Taylor expansion fails and bias cannot be controlled. No convergence rates can be derived.

---

10. Why is variance $\approx 1/(nh)$ in 1D?

About nh observations fall in a kernel window of width h. Averaging reduces variance as $1/(nh)$.

B. Kernel Functions & Equivalent Kernels

11. What conditions must a kernel satisfy?

Symmetric, integrates to 1, finite variance, square-integrable. Symmetry eliminates odd terms in Taylor expansions.

---

12. Why does bandwidth dominate kernel choice?

All reasonable kernels differ only by constants in asymptotic bias/variance.
h determines the *amount* of smoothing.

---

13. Why can local linear weights be negative?

To reproduce linear functions and correct boundary asymmetry, weighted least squares induces negative weights.

---

14. What is an equivalent kernel?

The implicit weighting function $W_{\text{eq}}(x)$ such that

$$\hat{m}(x) = \sum_i W_{\text{eq},i}(x) Y_i.$$

Shows how smoothers behave, including boundary effects.

---

15. How does local linear regression correct boundary bias?

It fits a line, not a constant. Asymmetric design near boundaries is compensated by asymmetric weights, removing first-order bias.

---

16. Why is Epanechnikov optimal but rarely used?

It minimizes the asymptotic MISE constant, but the improvement over Gaussian or rectangular kernels is negligible in practice.

---

17. How does local polynomial regression adapt to design density?

Weighted least squares automatically gives wider effective windows where $f_X(x)$ is small and narrower where $f_X(x)$ is large.

---

18. Why do some kernels create oscillations?

Higher-order kernels have negative lobes $\rightarrow$ ringing artifacts, especially near boundaries or when sample size is small.

---

19. Why is symmetry desirable in a kernel?

Symmetry ensures $\int uK(u)\,du = 0$, eliminating first-order bias terms.

---

20. Why are compact-support kernels sometimes preferred?

They make smoothing strictly local, reduce computational burden, and avoid heavy tails in weighting schemes.

C. Histogram vs KDE vs Local Polynomial Regression

21. Why does the histogram have MISE rate $n^{-2/3}$?

Bias $= O(h)$, variance $= O(1/(nh))$, balance $\to h \asymp n^{-1/3} \to$ MISE $n^{-2/3}$.

---

22. Why is the histogram discontinuous?

It is piecewise constant by construction, with abrupt jumps at bin boundaries.

---

23. Why does bin origin t_0 matter?

Bin boundaries determine membership; shifting t_0 can move observations between bins and drastically alter the estimate.

---

24. Why prefer KDE over histogram?

KDE uses smooth weights, no bin alignment issues, has smoother bias/variance behavior, and achieves faster $n^{-4/5}$ rate.

---

25. Why is local quadratic sometimes better than local linear?

It reduces curvature-driven bias in $m''(x)$ but increases variance; useful with large n and strong curvature.

---

26. Why is local constant regression less stable near boundaries?

It averages only the nearest points; with missing data on one side, imbalance creates strong one-sided bias.

---

27. Why does KDE avoid discontinuities?

It weighs observations smoothly by distance, so contributions change continuously with x.

---

28. Why do KDE and local linear share the same convergence rate $n^{-4/5}$?

Both have $O(h^2)$ bias and $O(1/(nh))$ variance. The polynomial order (0 vs 1) changes constants but not leading terms.

---

29. Why is local quadratic unstable at boundaries?

The design matrix becomes ill-conditioned due to one-sided data, magnifying variance.

---

30. What happens when h is too small or too large?

- Too small: undersmoothing $\rightarrow$ high variance, noisy curves.
- Too large: oversmoothing $\rightarrow$ high bias, loss of structure.

D. Bandwidth Selection

31. Why does LSCV undersmooth?

The CV score is noisy with many local minima. Minimizing it picks bandwidths that chase random bumps in the data.

---

32. What is the normal reference rule and when is it bad?

Assumes f is normal; for skewed or multimodal densities, curvature is misestimated, giving poor h.

---

33. What is the idea behind plug-in selectors?

Estimate unknown curvature terms like $R(f'') = \int (f''(x))^2 dx$ with a pilot estimator and plug into the formula for the optimal bandwidth.

---

34. Why undersmooth for confidence intervals?

At MSE-optimal h, bias and standard error are same order. Undersmoothing makes bias negligible relative to variance, validating normal approximations.

---

35. Why is multivariate bandwidth selection difficult?

You may need a full bandwidth matrix; rotations, scaling, and d-dimensional windows make CV/noise problems exponentially harder.

---

36. Why is pointwise MSE not enough for density estimation?

It does not capture global behavior. MISE integrates over the domain and is both analytically tractable and globally meaningful.

---

37. Why does LSCV fail badly in multimodal settings?

It tends to choose overly small h to reveal random "bumps," mistaking noise for genuine modes.

---

38. Why is cross-validation more stable in regression than in KDE?

Regression residuals provide direct error estimates; KDE relies on leave-one-out density estimates, which can be extremely unstable.

---

39. Why does smoothed cross-validation improve stability?

Smoothing the CV objective dampens noise, reducing erratic minima and leading to more reliable h.

---

40. Why might no single bandwidth work well across the entire domain?

Optimal local smoothness varies with curvature $m''(x)$ and design density $f_X(x)$. A global h cannot adapt simultaneously to flat and highly curved regions.

More Questions!

1. What does "nonparametric" really mean in this course? How is it different from parametric modeling?

A. "Nonparametric" means we do **not** commit to a fixed finite-dimensional functional form for the distribution or regression function. Instead, the target is often a **function** such as a density f, regression function m, or distribution function F.

Compared to parametric models:

- **Pro:** Flexible, avoids model misspecification.
- **Con:** Slower convergence, more tuning, more sensitive to dimension.

---

2. What is the fundamental bias–variance tradeoff in nonparametric smoothing?

A. Smoothing relies on averaging over local neighborhoods.

- **More smoothing (large h):** Lower variance, higher bias.
- **Less smoothing (small h):** Lower bias, higher variance.

Choosing h balances these forces.

---

3. What is the bandwidth, conceptually?

A. The bandwidth is the **size of the local neighborhood** used for estimation. When estimating at a point x, it answers:

> How far away am I allowed to look for data that inform my estimate at x?

It is the most important tuning parameter.

---

4. How would you explain a kernel density estimator (KDE) to someone who knows only histograms?

A. A KDE is a smoothed histogram:

- Each data point contributes a smooth bump (a kernel).
- The bandwidth determines bump width.
- Adding the bumps yields a smooth density estimate.

It avoids abrupt bin edges and produces a more interpretable shape.

---

5. Does kernel choice matter as much as bandwidth choice in KDE?

A. No. The kernel affects only small efficiency constants. Bandwidth determines the main bias–variance tradeoff.

In practice:

- Pick a standard kernel (e.g., Gaussian).
- Focus on selecting the bandwidth.

---

6. What features should you look for when interpreting a KDE plot?

A. Look for structure that persists across reasonable bandwidths:

- Number of modes
- Tail behavior
- Skewness
- Whether features are robust or likely noise

---

7. Conceptually, how is nonparametric regression different from linear regression?

A. Linear regression assumes a **global parametric form**. Nonparametric regression estimates the function $m(x)$ **locally**, allowing the shape to emerge from the data without assuming linearity.

---

8. Why is local linear regression often preferred over Nadaraya–Watson (NW)?

A. Three main reasons:

1. **Boundary correction:** NW has strong bias at boundaries.
2. **Automatic carpentry:** Local linear fits automatically correct first-order bias.
3. **Better equivalent kernels:** Weighting is more stable and less variable.

---

9. What is an equivalent kernel and why do we care?

A. An equivalent kernel expresses a local polynomial estimator as a **weighted average**, making it easier to analyze:

- Bias and variance
- Boundary behavior
- Comparisons among estimators

It clarifies how the estimator weights nearby points.

---

10. What is the curse of dimensionality in nonparametric estimation?

A. As dimension d increases:

- Data become sparse.
- Local neighborhoods contain fewer observations.
- Convergence rates deteriorate, e.g.,

$$\text{MISE} \sim n^{-4/(4+d)}.$$

Nonparametric estimators become ineffective in high dimensions unless additional structure is imposed.

---

11. How does sparsity manifest as dimension increases?

A. In high dimensions, points are typically:

- Far from each other
- Far from the center
- Isolated unless n is extremely large

Thus, smoothing either uses too few points (variance) or too wide a region (bias).

---

12. What is cross-validation doing when selecting a bandwidth?

A. CV approximates prediction error by:

- Leaving one observation out
- Predicting it from the remaining data
- Averaging squared errors over all observations

The bandwidth minimizing this approximates the best predictive fit.

---

13. Why does CV often produce "undersmoothed" bandwidths?

A. CV focuses on prediction, which sometimes rewards following noise. Thus, it tends to select bandwidths slightly too small for global MISE, producing wiggly estimates.

---

14. What is the empirical distribution function (EDF) and why is it central?

A. The EDF is

$$\hat{F}(x) = \frac{1}{n} \sum_{i=1}^{n} 1(X_i \leq x).$$

It is central because:

- It is the nonparametric MLE of F.
- It converges uniformly to F.
- It supports rank tests, KS tests, bootstrap resampling, and quantile estimation.

---

15. Why are rank-based tests considered "distribution-free"?

A. For continuous F, the **distribution of ranks does not depend on** F. Thus, rank-based tests have null distributions that are universal, making them robust to heavy tails and outliers.

---

16. What problem does deconvolution solve, and why is it difficult?

A. Deconvolution seeks to estimate the density of an unobserved variable X when we observe

$$Y = X + \varepsilon.$$

Difficulty: In the Fourier domain, we divide by $\phi_\varepsilon(t)$. When $\phi_\varepsilon(t)$ is small, noise is amplified. The problem is ill-posed and requires strong smoothing.

---

17. Why is the bootstrap especially useful in nonparametric settings?

A. Nonparametric estimators have complicated sampling distributions. The bootstrap:

- Estimates variability
- Provides confidence bands
- Requires minimal assumptions

It is a practical tool for inference with flexible estimators.

---

18. How do bootstrap confidence bands differ from parametric ones?

A. Bootstrap bands:

- Do not assume linearity or normality
- Capture nonlinear features of the estimator
- Can provide simultaneous (global) coverage
- Reflect empirical variability directly

They are especially useful for KDEs and local polynomial regressions.

---

19. How would you justify using a nonparametric method to a collaborator who prefers parametric models?

A. Emphasize:

- **Flexibility:** No assumption of functional form.
- **Model checking:** Compare nonparametric and parametric fits.
- **Interpretability:** The shape is visually clear.
- **Robustness:** Less sensitive to misspecification.

Nonparametric methods serve both diagnostic and primary roles.

---

20. What signs suggest a nonparametric regression or density estimate may be untrustworthy?

A. Warning signs include:

- Large changes under small bandwidth adjustments
- Strong boundary artifacts
- Excessive wiggliness
- Clear contradictions with raw data
- Implausible shapes not supported by subject-matter knowledge

Reliable features should persist across tuning choices.
