Nonparametric Techniques

Nonparametric Techniques

- w/o assuming any particular distribution
  - the underlying function may not be known (e.g. multi-modal densities)
  - too many parameters
- Estimating density distribution directly
- Transform into a lower-dimensional space where parametric techniques may apply (more on this later on dimension reduction)

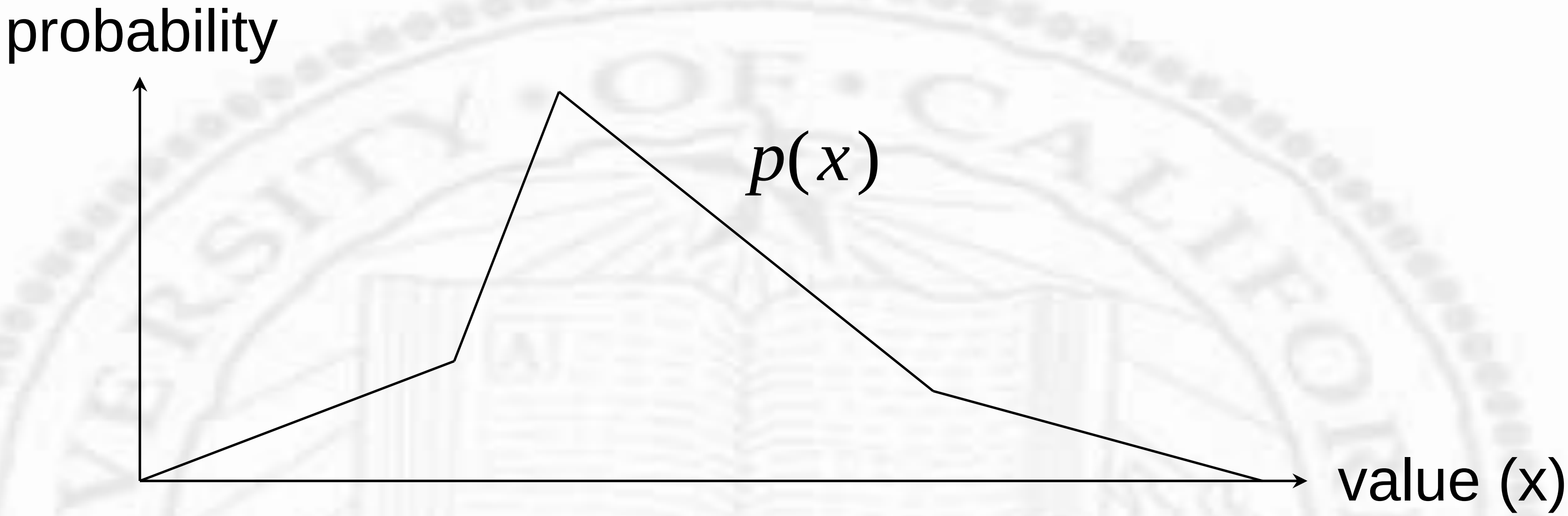
Example

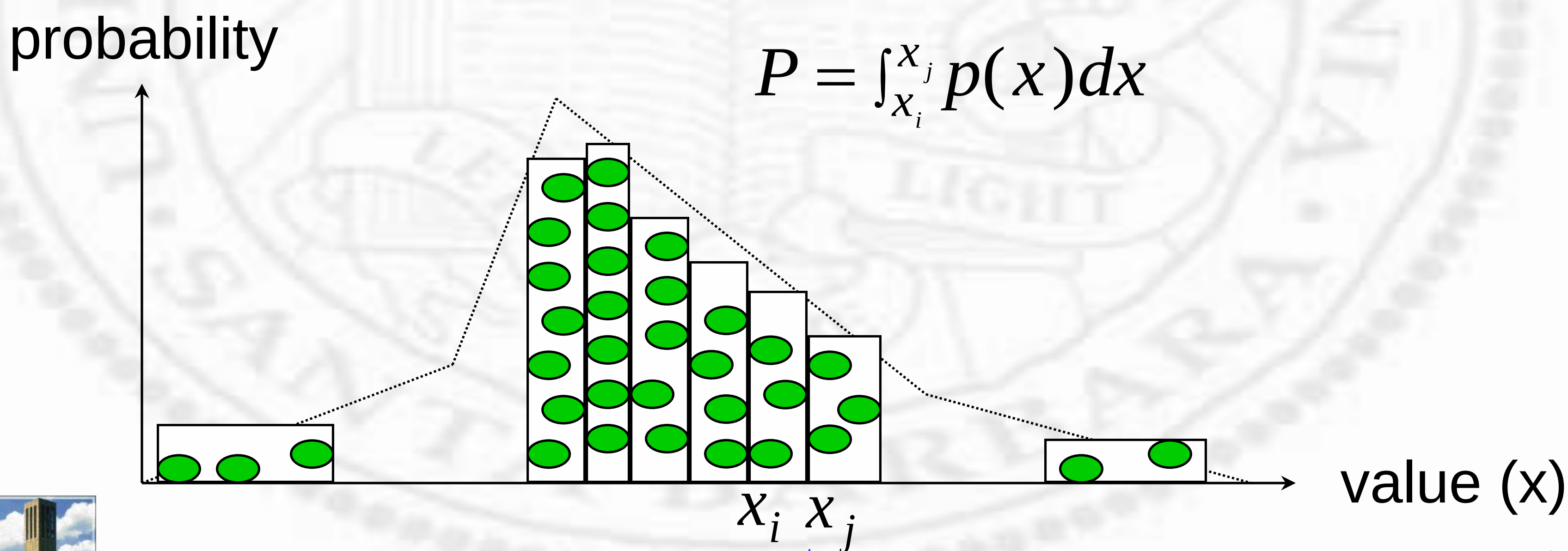
- Estimate the population growth, annual rainfall, etc. in the US
- p(x,y)dxdy is the probability of rain fall in [x,x+dx,y,y+dy]

Example (cont.)

- A simple parametric model for p(x,y) probably does not exist
- In stead
  - partition the area into a lattice
  - At each (x,y), count the amount of rain r(x,y)
  - Do that for a whole year
  - Normalize Σ r (x,y) = 1

Density estimation





Density estimation

- From equation

$$P = \int_{x_i}^{x_j} p(x)dx \cong p(x)(x_j - x_i)$$

- From observation

$$P = \frac{k}{n}$$

- Hence

$$p(x) \cong \frac{k/n}{(x_j - x_i)} = \frac{k/n}{V}$$

Comparison

- In Reality:
- The number of training samples is limited
- if V is too small, k becomes erratic
  - What does 0 mean?
- if V is too large, $p(x)$ is not representative

- In theory:
- If *n* becomes infinitely large, *k/n* approaches the probability, *p(x)* = *(k/n)/V* is then only a space average
- Hence, *V* must be allowed to go to zero as *n* goes to infinity

$$p(x) \cong \frac{k/n}{(x_j - x_i)} = \frac{k/n}{V}$$

In Theory

- Theoretically, we can use a sequence of samples with increasing size for estimation
- Then

$$p_n(x) \rightarrow p(x) \quad if$$
$$(1) \lim_{n->\infty} V_n = 0$$
$$(2) \lim_{n->\infty} k_n = \infty$$
$$(3) \lim_{n->\infty} \frac{k_n}{n} = 0$$

Two different approaches

- Constrain the region size
  - Shrink the region to maintain good locality (Parzen Windows)
- Constrain the sample size
  - Enlarge the number of samples to maintain good resolution (K_n-nearest-neighbors)

Parzen Windows

- Use a windowing function, e.g.
- A sequence of *n* regions can be defined

$$\phi(x) = \begin{cases} 1 & |x| \le \frac{1}{2} \\ 0 & otherwise \end{cases} \quad or \quad \frac{1}{2\pi} e^{-\frac{x^2}{2}}$$

$$\phi_n(x) = \phi(x / h_n)$$
$$h_n = \frac{h_1}{\sqrt{n}}$$

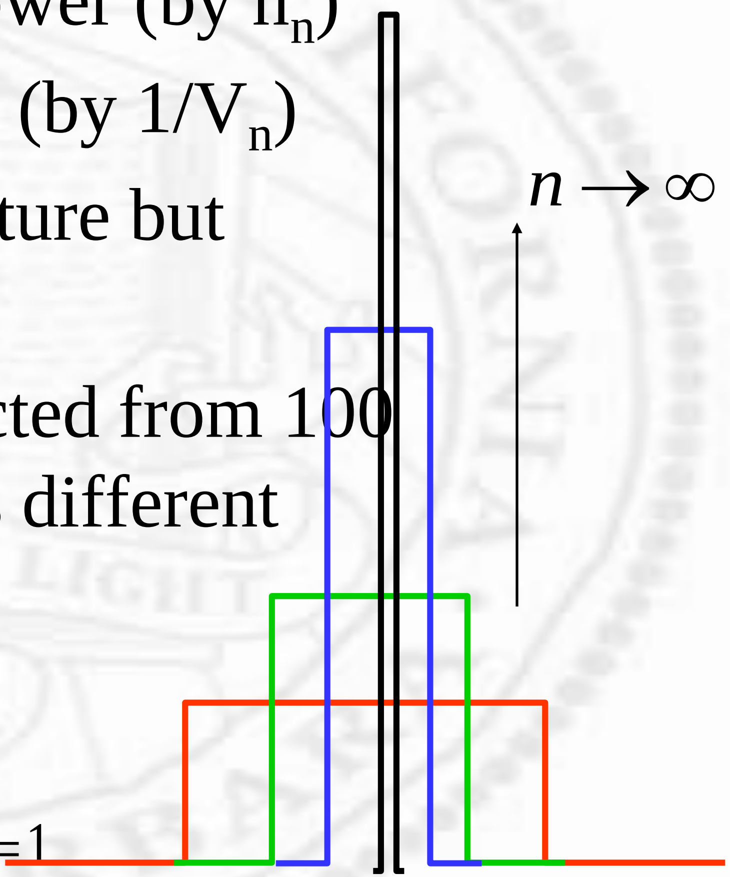
$$k_n = \sum_{i=1}^{n} \phi_n(x - x_i) = \sum_{i=1}^{n} \phi(\frac{x - x_i}{h_n})$$

$$p_n(x) = \frac{1}{n} \sum_{i=1}^{n} \underbrace{\frac{1}{V_n} \phi(\frac{x - x_i}{h_n})}_{} = \frac{1}{n} \sum_{i=1}^{n} \underbrace{\delta_n(x - x_i)}_{\text{By definition}}$$

By definition

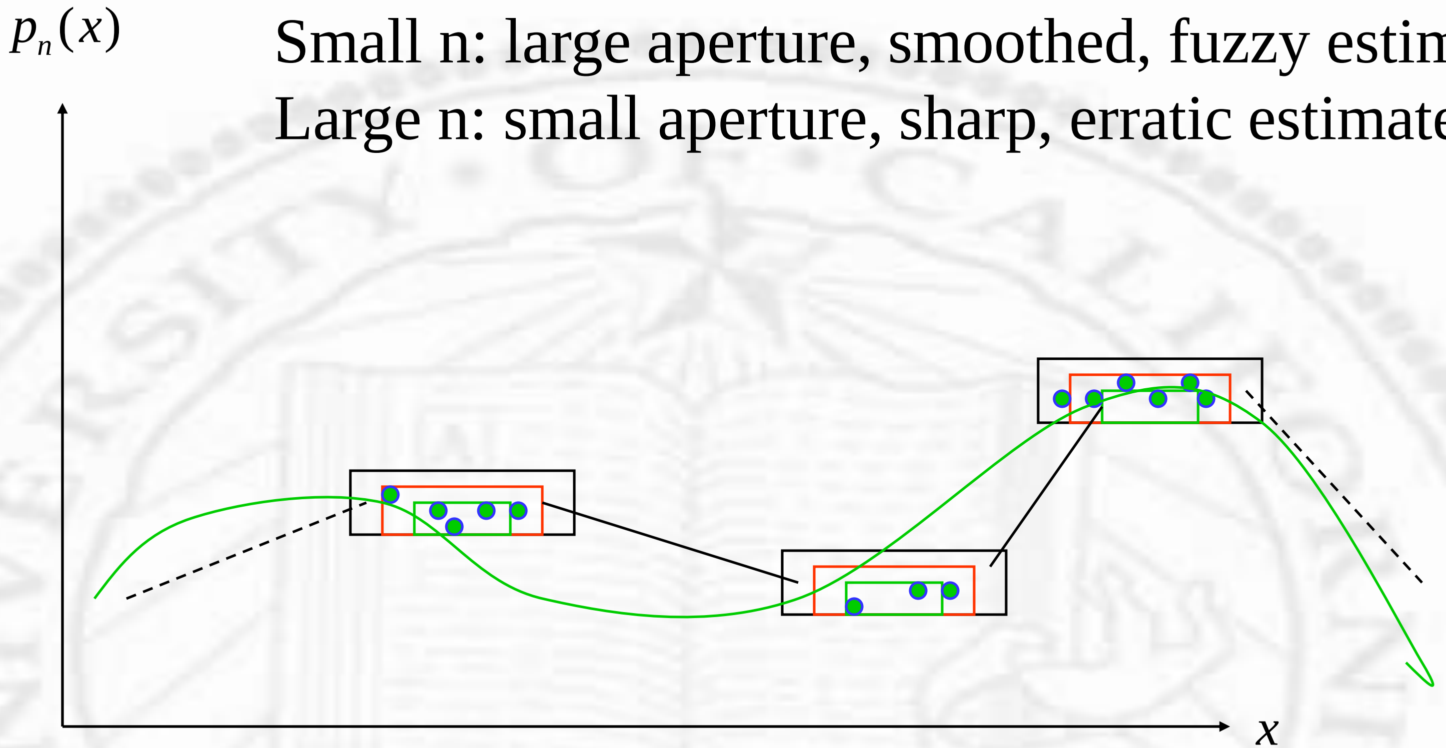
Parzen Window (cont.)

- As n increases
  - The window becomes narrower (by h_n)
  - The window becomes taller (by $1/V_n$)
  - Sampling with smaller aperture but higher focus
  - The same 100 dollars collected from 100 people and from 1 person is different (per person)



$$\int \delta_n(x - x_i)dx = \int \frac{1}{V_n}\phi\left(\frac{x - x_i}{h_n}\right)dx = \int \phi(u)du = 1$$

Small n: large aperture, smoothed, fuzzy estimate
Large n: small aperture, sharp, erratic estimate



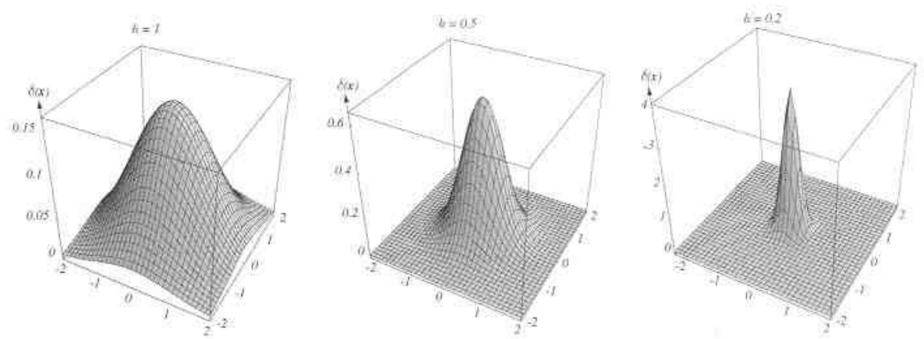
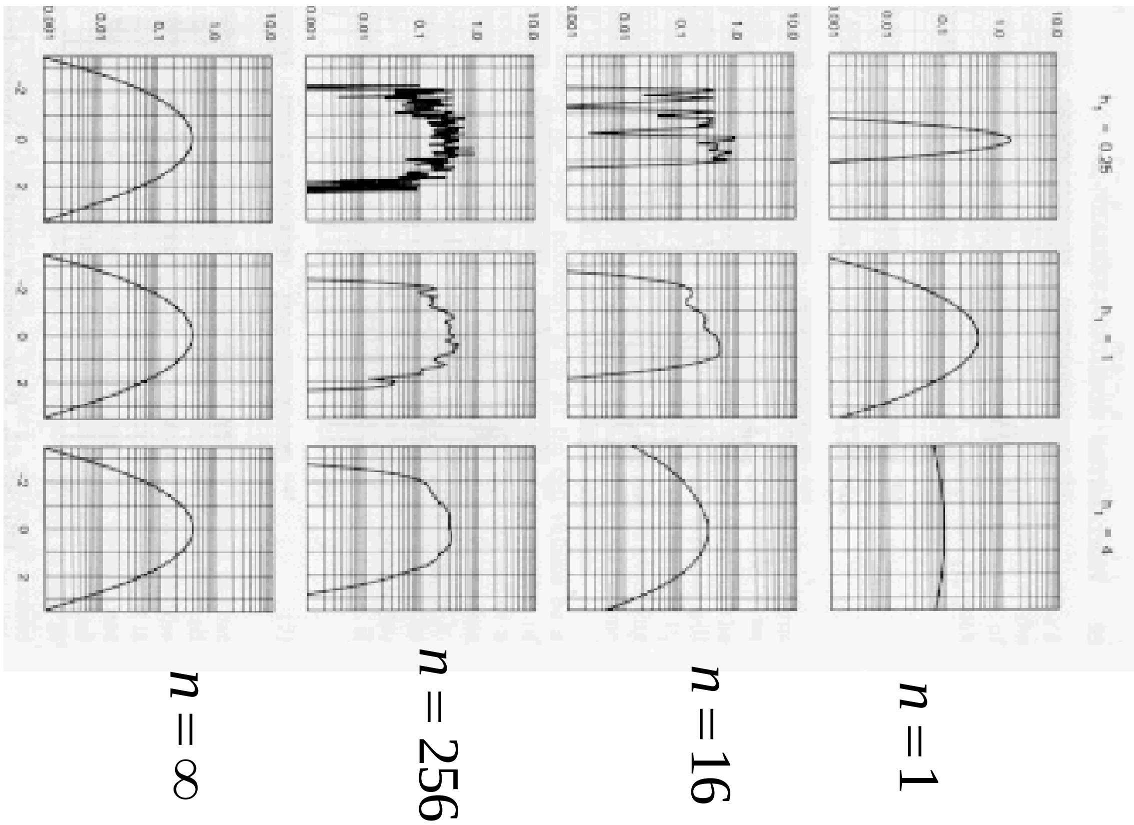


FIGURE 4.3. Examples of two-dimensional circularly symmetric normal Parzen windows for three different values of h. Note that because the $\delta(\mathbf{x})$ are normalized, different vertical scales must be used to show their structure.

Examples of the Parzen Window Estimation

Example I: $\varphi()$ as a Gaussian window of various width.



2D Sampling

- Five samples
- Windowing func:

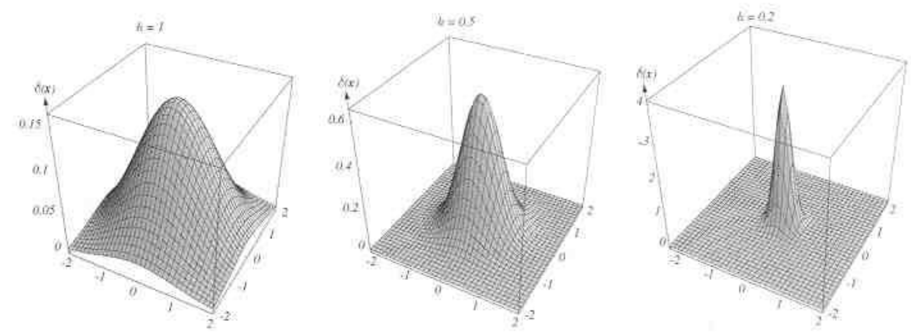
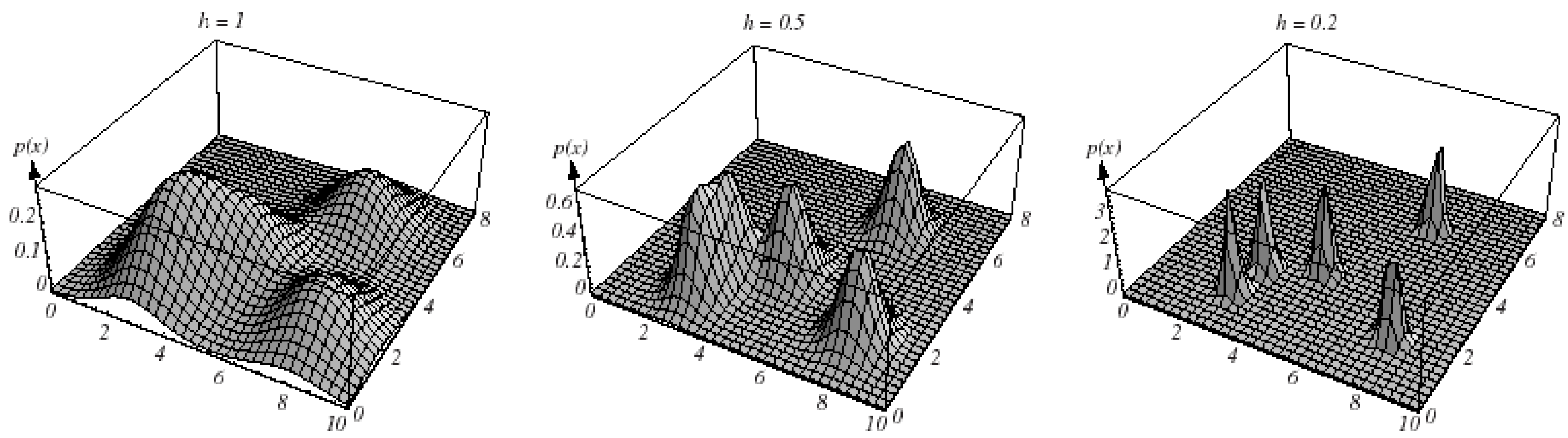


FIGURE 4.3. Examples of two-dimensional circularly symmetric normal Parzen windows for three different values of h. Note that because the $\delta(\mathbf{x})$ are normalized, different vertical scales must be used to show their structure.



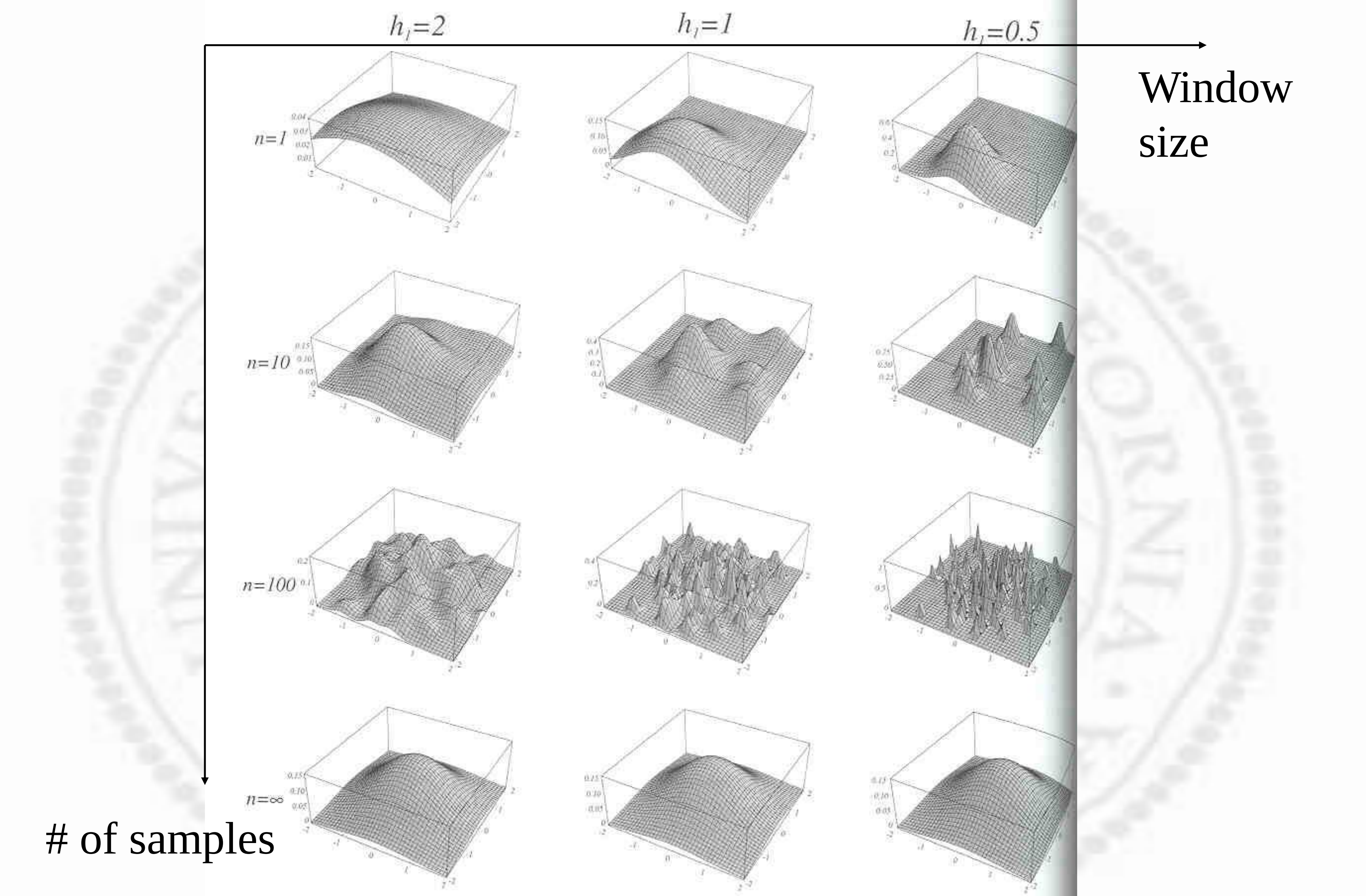
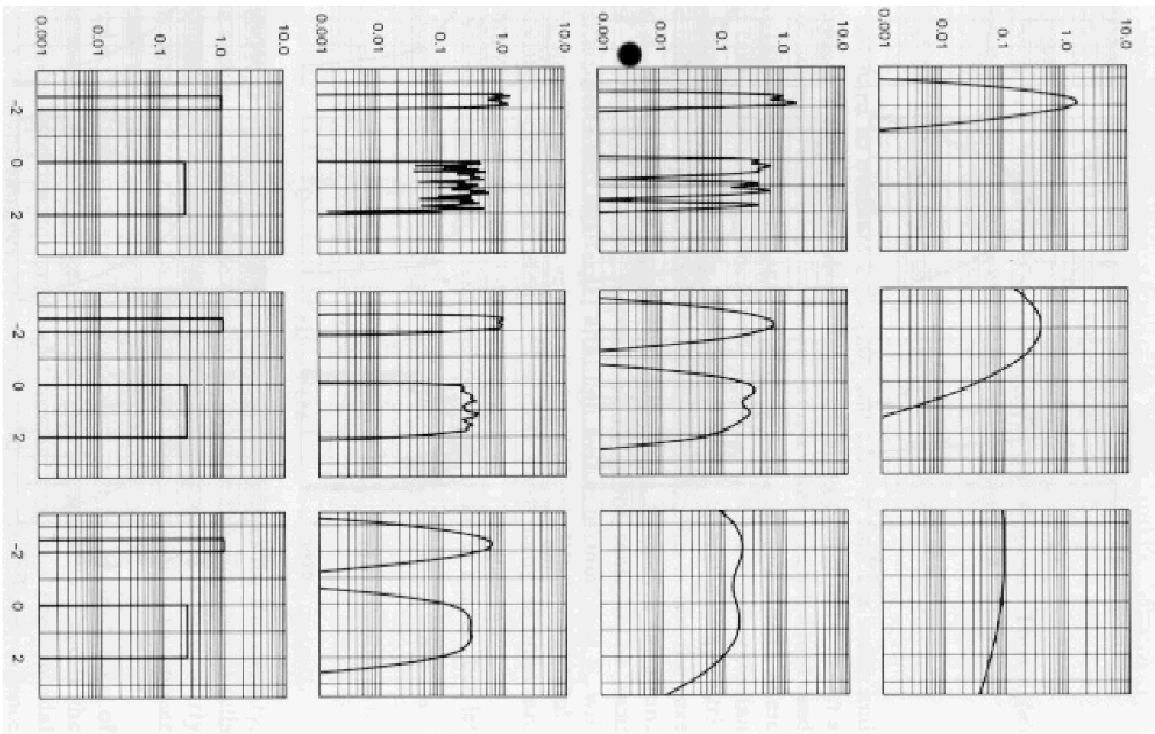


FIGURE 4.6. Parzen-window estimates of a bivariate normal density using different window widths and numbers of samples. The vertical axes have been scaled to best show the structure in each graph. Note particularly that the $n = \infty$ estimates are the same (and match the true distribution), regardless of window width.

Examples of the Parzen Window Estimation -continue

Example II: $\varphi()$ as a Gaussian window of various width.



Does it work?

- "*Work*" in the sense that you if you are able to shrink down the window size as much as you want (certainly, you must simultaneously increase the number of samples available), then the limit of the profile should be the correct probability
- This implies (treating p_n as a random variable)
  - $E(p_n(\mathbf{x}))=p(\mathbf{x})$
  - $Var(p_n(\mathbf{x})) \to 0$

Convergence of Mean

- Will p_n ($\mathbf{x}$) goes to p($\mathbf{x}$)?
  - If *n* goes to infinity
    - $\mathbf{x}_i$ will cover all possible $\mathbf{x}$ (summation to integration)
    - with p($\mathbf{x}$) distribution (weighted by p($\mathbf{x}$))

$$
\begin{aligned}
\bar{p}_n(\mathbf{x}) &= E[p_n(\mathbf{x})] \\
&= \frac{1}{n}\sum_{i=1}^{n} E[\frac{1}{V_n}\varphi(\frac{\mathbf{x}-\boxed{\mathbf{x}_i}}{h_n})] \\
&= \int \frac{1}{V_n}\varphi(\frac{\mathbf{x}-\boxed{\mathbf{v}}}{h_n}) p(\mathbf{v}) d\mathbf{v} \\
&= \int \delta_n(\mathbf{x}-\mathbf{v}) p(\mathbf{v}) d\mathbf{v} = p(\mathbf{x})
\end{aligned}
$$

Sample **v** appears with probability p(**v**)

Convergence of Variance

- Will $p_n(\mathbf{x})$ always end up at $p(\mathbf{x})$ for certain?
  - nV_n must approach infinity, even V_n when goes to zero

$$\sigma_n^2(\mathbf{x}) = \sum_{i=1}^{n} E\left[\left(\frac{1}{nV_n}\phi(\frac{\mathbf{x}-\mathbf{x}_i}{h_n}) - \frac{1}{n}\bar{p}_n(\mathbf{x})\right)^2\right]$$

$$= \sum_{i=1}^{n} nE\left[\left(\frac{1}{n^2V_n^2}\phi^2(\frac{\mathbf{x}-\mathbf{x}_i}{h_n})\right)\right] - \frac{1}{n}\bar{p}_n^2(\mathbf{x})$$

$$= \frac{1}{nV_n}\int \frac{1}{V_n}\phi^2(\frac{\mathbf{x}-\mathbf{v}}{h_n})p(\mathbf{v})d\mathbf{v} - \boxed{\frac{1}{n}\bar{p}_n^2(\mathbf{x})}$$ -> 0 as n-> infinity

$$\leq \frac{1}{nV_n}\sup(\phi(\cdot))\int \frac{1}{V_n}\phi(\frac{\mathbf{x}-\mathbf{v}}{h_n})p(\mathbf{v})d\mathbf{v}$$

$$\sigma_n^2(\mathbf{x}) \leq \frac{\sup(\phi(\cdot))\bar{p}_n(\mathbf{x})}{nV_n}$$

k_n-nearest-neighbor

- Parzen window size hard to estimate
- Constrain the number of data items instead of the size of the window
- $k_n = \sqrt{n}$ enlarge window around **x** to enclose that many samples, then

$$p_n(x) = \frac{k_n / n}{V_n}$$

k_n-nearest-neighbor

- Intuitively, as *n* increases
  - k_n should increase (for good representation)
  - V_n should decrease (for good localization)
  - The following conditions guarantee convergence

$$\lim_{n \to \infty} k_n = \infty$$

$$\lim_{n \to \infty} \frac{k_n}{n} = 0$$

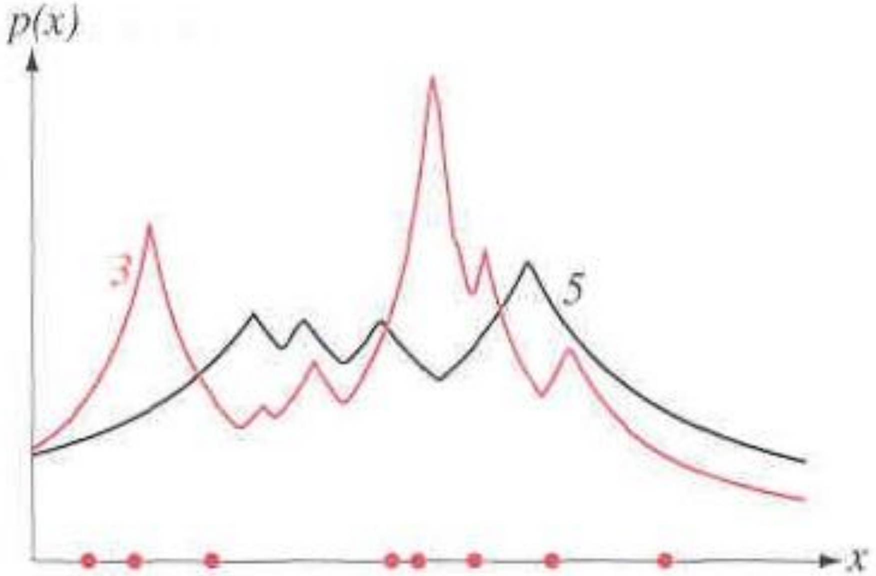
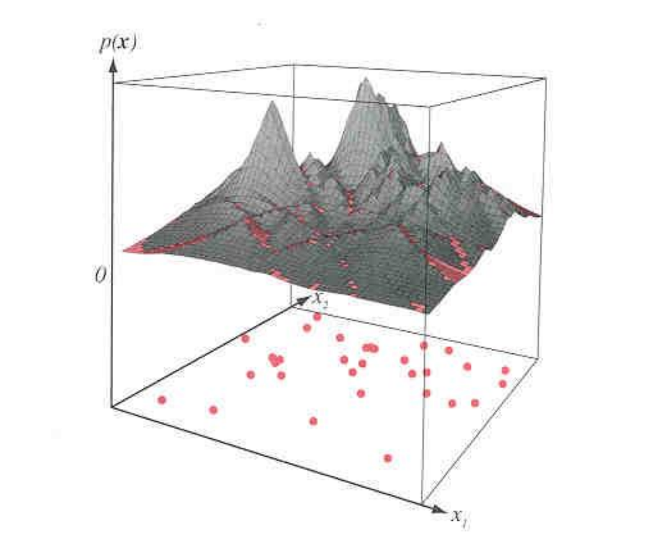


FIGURE 4.10. Eight points in one dimension and the *k*-nearest-neighbor density estimates, for $k = 3$ and 5. Note especially that the discontinuities in the slopes in the estimates generally lie *away* from the positions of the prototype points.

Sharp spikes around data points:
Kn=1, the probability estimate is infinity at data point (region size is zero to capture 1 sample)



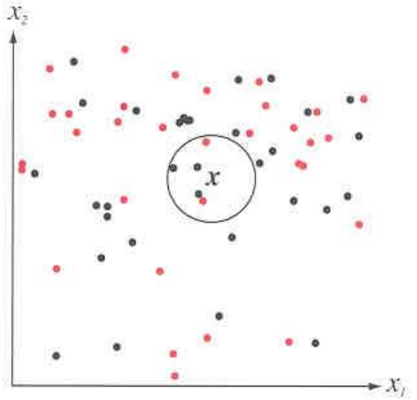
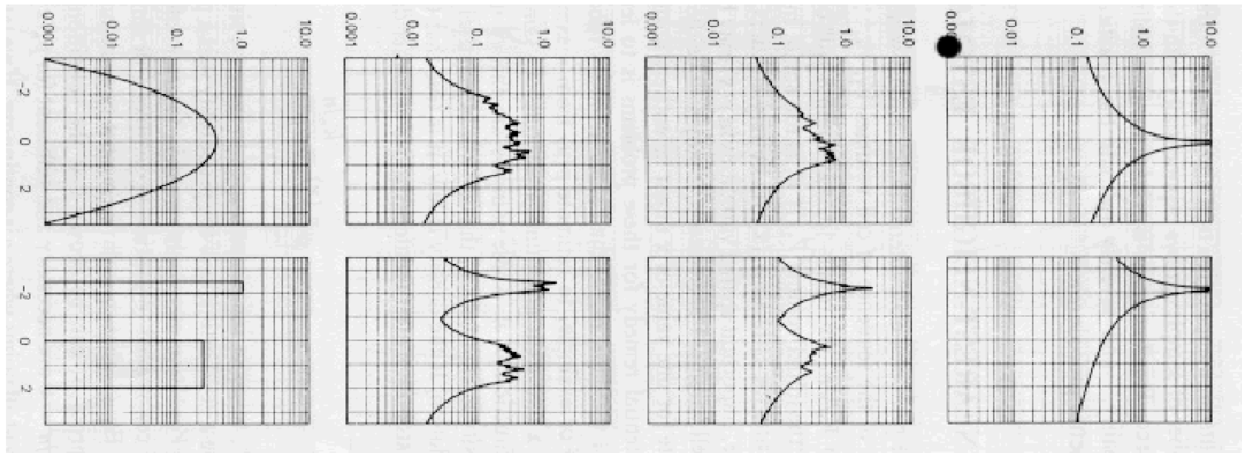


FIGURE 4.15. The k-nearest-neighbor query starts at the test point **x** and grows a spherical region until it encloses k training samples, and it labels the test point by a majority vote of these samples. In this $k = 5$ case, the test point **x** would be labeled the category of the black points.

Examples of the k_n-Nearest-Neighbor Method



$n = \infty$	$n = 256$	$n = 16$	$n = 1$
$k_n = \infty$	$k_n = 16$	$k_n = 4$	$k_n = 1$

An Example

- Estimating $p(\varpi_i \mid \mathbf{x})$
  - n tagged samples
  - a volume V around **x** captures k samples, k_i of them are ϖ_i

$$p_n(\mathbf{x}, \varpi_i) = \frac{k_i / n}{V}$$

$$p_n(\varpi_i \mid \mathbf{x}) = \frac{p_n(\mathbf{x}, \varpi_i)}{\sum_{j=1}^{c} p_n(\mathbf{x}, \varpi_j)} = \frac{\frac{k_i / n}{V}}{\sum_{j=1}^{c} \frac{k_i / n}{V}} = \frac{\frac{k_i / n}{V}}{\frac{k / n}{V}} = \frac{k_i}{k}$$

Comparison

- Parametric
  - simple and analytical
  - may not fit well real-world densities
- Non-parametric
  - flexible and fit all densities
  - need to remember all samples

One Final Note

- Here we talk about Parzen window and k_n-nearest-neighbor rule as a way to estimate *a single* probability density
- This rule is equally useful at labeling a sample against multiple probable classes (densities)
- More on that in linear discriminant function

More Realistic Scenarios

- Drake's Equation
  - Rate of start formation, fraction of stars having planets, average # of planets per star that support life, fraction of such stars actually develop life, fractions of such stars actually develop civilization, such civilization have communication, length of time such civilization actually release signals

More Realistic Scenarios

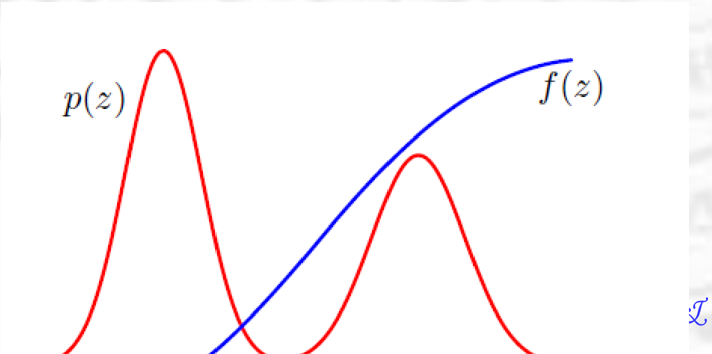
- Chance of a person develops cancer (ancestry, birth place, how raised, living habits, education history, work history, exercise habit, income, debt, food intake, etc.)
- Chance of a person contributes to political campaign (…)

Curse of Dimensionality

- Not possible to estimate distributions in such high-dimensional space
- # of samples needed are generally infinitely large

Practical Usage

- X = rand(3,3)
- Sampling based on certain distribution (default is uniform)
- Need to evaluate certain expectation
- Technology advances by alien contact
- Life expectancy (for cancer case)
- Amount of money for political campaigns



$$\mathbb{E}[f] = \int f(\mathbf{z})p(\mathbf{z})\,\mathrm{d}\mathbf{z}$$

General Idea

- Finite number samples: sample mean/variance to estimate population mean/variance
  - $z^{(l)}$, $l = 1, \ldots, L$

$$\widehat{f} = \frac{1}{L}\sum_{l=1}^{L} f(\mathbf{z}^{(l)}).$$

$$\mathrm{var}[\widehat{f}] = \frac{1}{L}\mathbb{E}\left[(f - \mathbb{E}[f])^2\right]$$

  - Samples may not be independent
  - Some distribution (uniform) is easier to sample than others
  - f(**z**) is small in regions where p(**z**) is large and vice versa

From One to Another

$$p(y) = p(z)\left|\frac{dz}{dy}\right|$$

$$p(y) = \lambda \exp(-\lambda y)$$

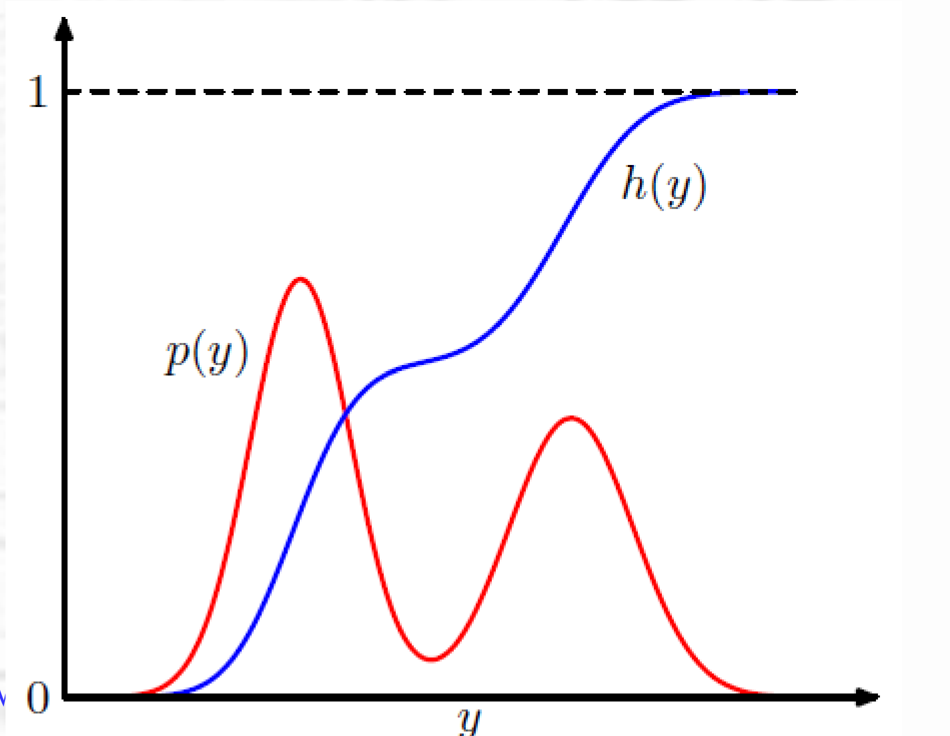
$$z = h(y) \equiv \int_{-\infty}^{y} p(\widehat{y})\,\mathrm{d}\widehat{y}$$

$$h(y) = 1 - \exp(-\lambda y)$$

$$y = h^{-1}(z)$$

$$y = -\lambda^{-1}\ln(1-z).$$

z: uniform
y: any known distribution
Sample z uniformly ==
Sample y based on p(y)



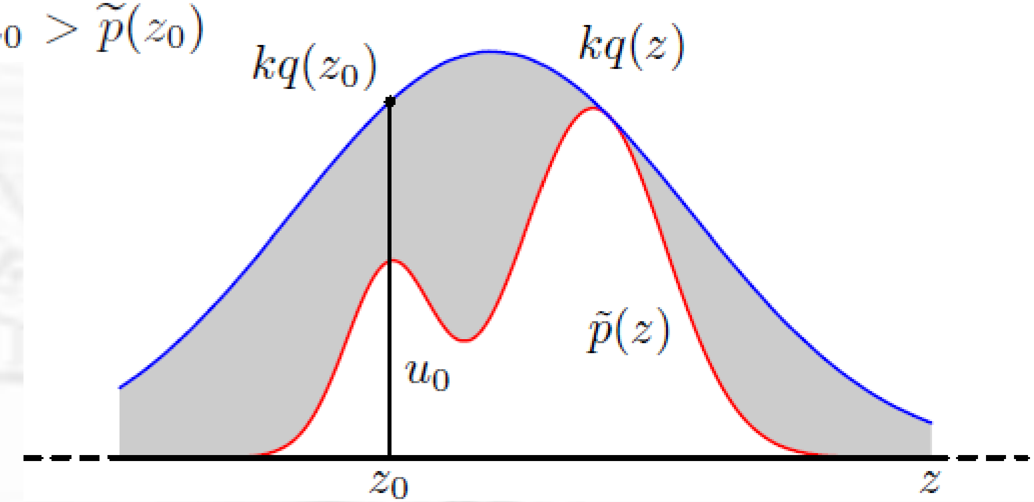
Multi-Dimensional

- Much more difficult
- Do not know the form
- Cannot get enough samples to populate the landscape
- How to generate IID samples?

Rejection Sampling

- A real distribution p(z) $$p(z) = \frac{1}{Z_p}\widetilde{p}(z)$$
- A proposal distribution q(z) $$kq(z) \geqslant \widetilde{p}(z)$$
- Procedure
  - Generate z_o from q(z)
  - Generate u_o from $[0, kq(z_o)]$ uniformly
  - Reject sample if $u_0 > \widetilde{p}(z_0)$
  - Otherwise, accept

$$\begin{aligned} p(\text{accept}) &= \int \{\widetilde{p}(z)/kq(z)\}\, q(z)\, \mathrm{d}z \\ &= \frac{1}{k}\int \widetilde{p}(z)\, \mathrm{d}z. \end{aligned}$$



Importance Sampling

- A real distribution p(z)
- A proposal distribution q(z)
- Procedure
  - Generate z_o from q(z), nothing rejected
  - $p(z^{(l)})/q(z^{(l)}))$: importance weight to account for sampling from wrong distribution

$$
\begin{aligned}
\mathbb{E}[f] &= \int f(\mathbf{z})p(\mathbf{z})\,\mathrm{d}\mathbf{z} \\
&= \int f(\mathbf{z})\frac{p(\mathbf{z})}{q(\mathbf{z})}q(\mathbf{z})\,\mathrm{d}\mathbf{z} \\
&\simeq \frac{1}{L}\sum_{l=1}^{L}\frac{p(\mathbf{z}^{(l)})}{q(\mathbf{z}^{(l)})}f(\mathbf{z}^{(l)}).
\end{aligned}
$$

MCMC

- Imagine
  - A very high-dimensional space
  - Samples occupy low-dimensional manifold in such a high-dimensional space
  - Choose a random start point
  - Wander about in the space, seeking out places with sample
  - With right "seek" strategy, samples generated along the walk have the right population characteristics

MCMC

- Successive sampling points are NOT independent, but form a Markov chain $q(\mathbf{z}|\mathbf{z}^{(\tau)})$
- Z* is generated at each step, accepted if probability > preset threshold $A(\mathbf{z}^{\star}, \mathbf{z}^{(\tau)}) = \min\left(1, \frac{\widetilde{p}(\mathbf{z}^{\star})}{\widetilde{p}(\mathbf{z}^{(\tau)})}\right)$
- Can be shown that the distribution of $z^{(\tau)}$ tends to p(z) as τ -> infinity
- So distribution of steps z's after some initial steps can be used to approximate p(z)
- For Metropolis algorithm, q has to be symmetrical q(a|b)=q(b|a)

Meropolis - Hastings

- f(x): proportional to p(x) – target distribution
- Given:
  - x_o: first sample
  - Q(x'|x): Markov process to generate next sample (x') given current sample (x), Q must be symmetrical (e.g., Gaussian)
- Iteration:
  - X' picking from Q(x'|x)
  - r=f(x')/f(x) >=1 accept, otherwise accept with prob r. If rejected, x'=x

Intuition

- A random walk model
  - Move into more likely region with prob 1
  - Move into less likely region with prob $\propto$likelihood
  - Stay in the high-density region of p(x)
- Caveats:
  - Samples are correlated
    - Discard initial samples
    - Take 1 out of n-th samples
  - Slow mixing for high-dimensional data (Gibbs sampling is better)

Gibbs Sampling

- Special case of MCMC Metropolis-Hastings
- From $x^{(i)}$ to $x^{(i+1)}$ by component-wide sampling, j-th variable in $x^{(i+1)}$ depends on
  - 1 to j-1 in (i+1)-th iterations
  - j+1 to n in (i)-th iteration

$$p\left(x_j^{(i+1)} | x_1^{(i+1)}, \ldots, x_{j-1}^{(i+1)}, x_{j+1}^{(i)}, \ldots, x_n^{(i)}\right)$$

Slice Sampling

- Random walk under the probability curve
- Start from an x_o with $f(x)>0$
- Randomly select height y, $0<y<=f(x)$
- Randomly select x’ lie within the slice, repeat