Priors on Images

Suyash P. Awate

PATTERN RECOGNITION AND MACHINE LEARNING

CHRISTOPHER M. BISHOP

APR Advances in Pattern Recognition

Markov Random Field Modeling in Image Analysis

Stan Z. Li

Third Edition

Springer

Bayes Rule

- For events A and B,
  $P(A|B) = P(B|A)\,P(A) / P(B)$
  - Follows from the definition of conditional probability

Thomas Bayes

- Statistician, philosopher, Presbyterian minister

- Approximate Bayesian computation
- Bayes error rate
- Bayes estimator
- Bayes factor
- Bayes linear statistics
- Bayes prior
- Bayes' rule
- Bayes' theorem
- Empirical Bayes method
- Evidence under Bayes theorem
- Hierarchical Bayes model
- Laplace–Bayes estimator
- Naive Bayes classifier
- Random naive Bayes

- Bayesian average
- Bayesian approaches to brain function
- Bayesian econometrics
- Bayesian efficiency
- Bayesian experimental design
- Bayesian Filtering Library
- Bayesian game
- Bayesian inference
- Bayesian inference in phylogeny
- Bayesian information criterion
- Bayesian linear regression
- Bayesian model selection
- Bayesian multivariate linear regression
- Bayesian network
- Bayesian poisoning
- Bayesian probability
- Bayesian Program Learning
- Bayesian search theory
- Bayesian spam filtering
- Bayesian statistics
- Bayesian tool for methylation analysis
- Dynamic Bayesian network
- International Society for Bayesian Analysis
- Recursive Bayesian estimation
- Robust Bayesian analysis
- Variable-order Bayesian network
- Variational Bayesian methods

Thomas Bayes

- Sir Harold Jeffreys (statistician) wrote that Bayes' theorem "is to the theory of probability what the Pythagorean theorem is to geometry"
- Bayes never published what would eventually become his most famous accomplishment; his notes were edited and published after his death by Richard Price

Bayesian Statistics

Likelihood

How probable is the evidence given that our hypothesis is true?

Prior

How probable was our hypothesis *before* observing the evidence?

$$\mathrm{P}(H \mid e) = \frac{\mathrm{P}(e \mid H)\,\mathrm{P}(H)}{\mathrm{P}(e)}$$

Posterior

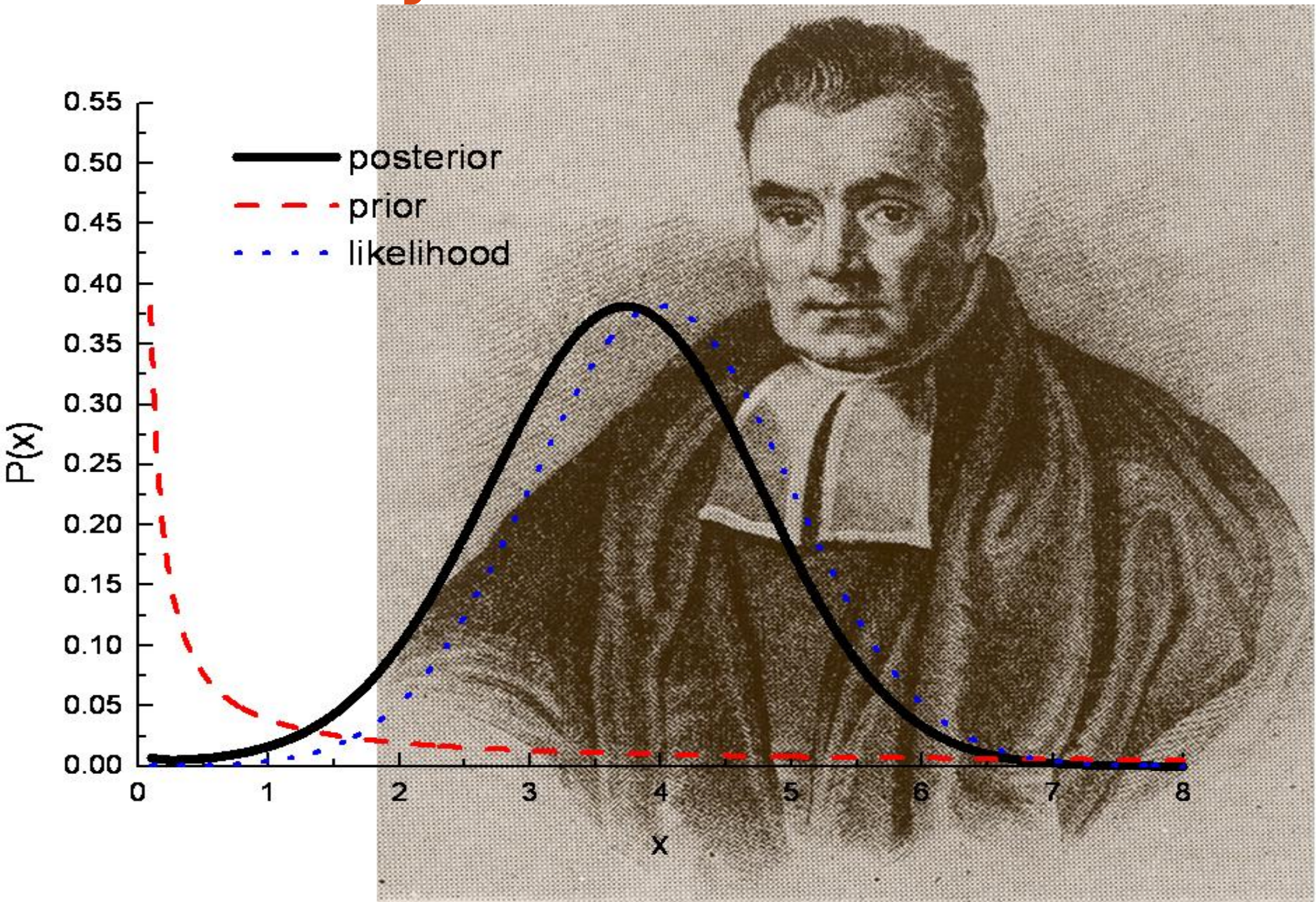
How probable is our hypothesis given the observed evidence? (Not directly computable)

Marginal

How probable is the new evidence under all possible hypotheses?

$$\mathrm{P}(e) = \sum \mathrm{P}(e \mid H_i)\,\mathrm{P}(H_i)$$

Bayesian Statistics

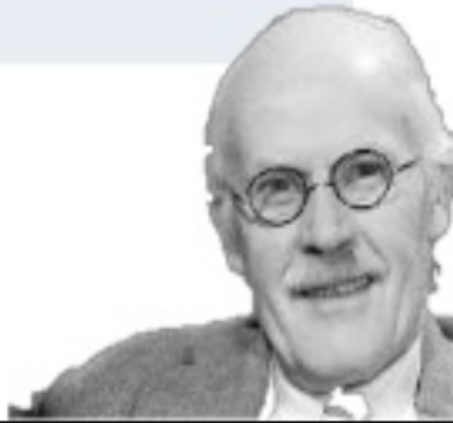


Bayesian Statistics

Frequentist	Bayesian
Probability is a long-run average	Probability is a degree of belief
There is a true Model, the Data is a random realization	The Data is true/fixed, Models have probabilities
Probability of the data given a hypothesis (Likelihood)	Probability of a hypothesis given the data
Each repeated experiment/observation starts from ignorance	Can incorporate prior knowledge: probabilities can be updated

Jerzy Neyman

Harold Jeffereys



Bayes Rule

- Bayes Rule
  - Application 1: Restoration
    - P (UncorruptedImage | CorruptedImage)
    - P (UncorruptedImage | CorruptedImage)
      = P (CorruptedImage | UncorruptedImage)
      * **P (UncorruptedImage)**
      / P (CorruptedImage)
    - Likelihood PDF = P (CorruptedImage | UncorruptedImage)
      - Corruption model
    - Prior PDF = **P (UncorruptedImage)**
      - **Our prior beliefs about the uncorrupted image**
    - Posterior PDF = P (UncorruptedImage | CorruptedImage)
      - Product of likelihood and prior (with normalizing factor)

Bayes Rule

- Bayes Rule
  - Application 2: Segmentation / Labeling
    - P (LabelImage | CorruptedImage)
    - P (LabelImage | CorruptedImage)
      = P (CorruptedImage | LabelImage)
      * **P (LabelImage)**
      / P (CorruptedImage)

Bayes Rule

- Simple Example (Gaussian)
  - Given
    - Data { x1, x2, …, xN }
    - Derived from a Gaussian distribution
    - *Known* std. dev. σ
    - *Unknown* mean μ
  - **Prior belief on μ**
    - **μ is derived from a Gaussian with mean μ0, std. dev. σ0**
  - Goal: Estimate μ, given data + prior
  - Strategy: Optimize μ to maximize posterior
    - **Maximum-a-posteriori (MAP) estimation**

Bayes Rule

- Simple Example (Gaussian)
  - What if we ignore the prior ?
    - ML estimation seen before
  - Assume sample mean = $\bar{u}$
  - Then, MAP estimate for μ is: $\mu = \frac{\sigma_0^2 \bar{u} + \sigma^2 \mu_0 / N}{\sigma_0^2 + \sigma^2 / N}$

Bayes Rule

- Simple Example (Gaussian)
  - What if we ignore the prior ?
    - ML estimation seen before
  - Assume sample mean = $\bar{u}$
  - Then, MAP estimate for μ is: $\mu = \frac{\sigma_0^2 \bar{u} + \sigma^2 \mu_0 / N}{\sigma_0^2 + \sigma^2 / N}$
  - Interpretation
    - What if $N = 1$?

      What if $N \to \infty$? (data dominates the prior)

      What if $\sigma_0 \to \infty$? (weak prior: ignore the prior)

      What if $\sigma_0 \to 0$? (strong prior: ignore the data)

Image Prior

- Space of images
  - Dimensions = N = number of voxels
  - Integer values (label images or data)
  - Real values (data)
- Prior beliefs on **uncorrupted** images
  - Based on physical / biological assumptions on objects being imaged

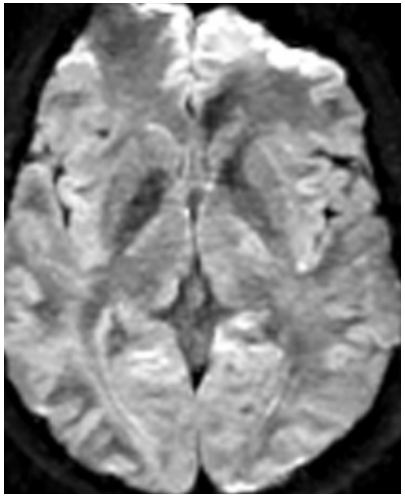
    1/3) Image intensities are spatially (piecewise) smooth

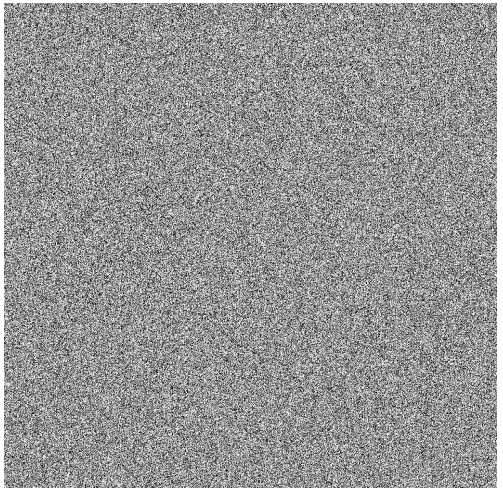
    2/3) Discontinuities / large changes possible only at object boundaries

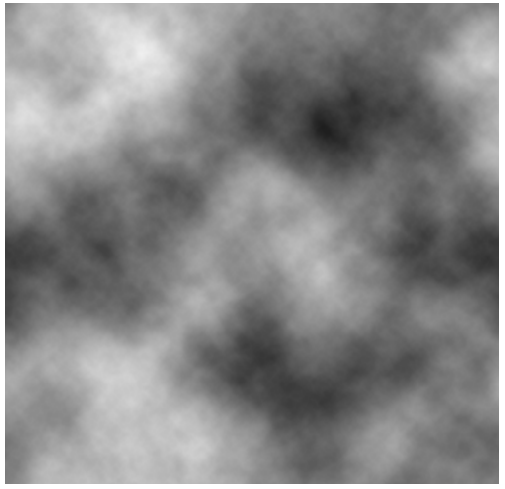
    3/3) Number of objects << number of pixels

Image Prior

- Assign a probability of occurrence to every image within the image space







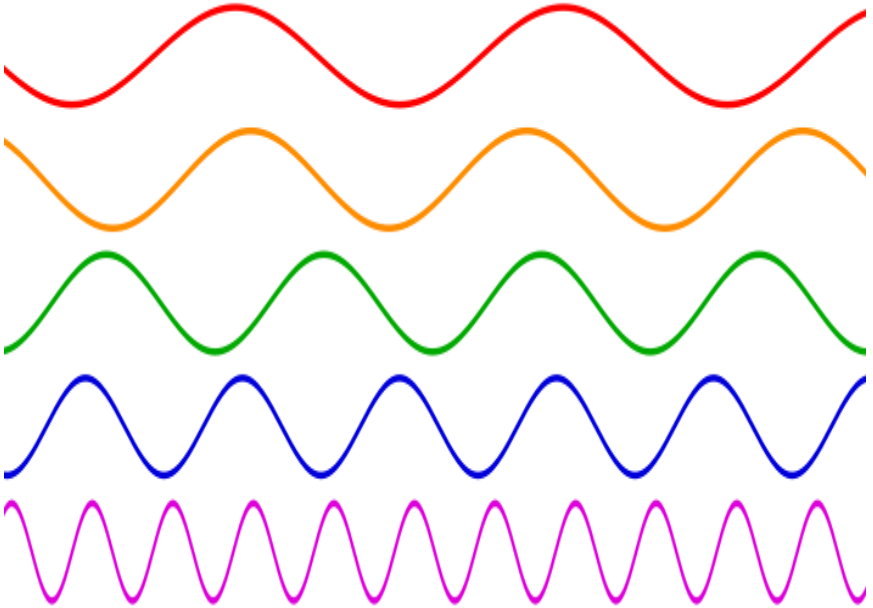


Image Prior

- Assign a probability of occurrence to every image within the image space
  - What if data = image in row 1 col 2 ?
  - What if data = image in row 2 col 1 ?

Image Prior

- Random field …
- Markov random field …
- Gibbs random field ...

Random Field

- **Topological space** =
  - A set of points (or sites) along with a set of neighborhoods for each point
  - e.g., for nD images, set S is, typically, a set of points on a N-dimensional Cartesian grid
- **Random field** =
  - Given a probability space {Ω, F, P},
    a random field is a collection of random variables $\{ X_i \}$
    indexed by values i in a topological space S
- Observed image x =
  single realization of random field $X = \{ X_i : \forall i \in S \}$
  under some PDF P(X)

Random Field

- **Neighborhood system** =
  - $N = \{ N_i \mid \forall i \in S \}$

    where N_i = set of sites neighboring site i
- **Neighbor relation** =

  (1) A site isn't a neighbor to itself: i not in N_i

  (2) Neighborhood relationship is mutual: $i \in N_j \Rightarrow j \in N_i$
- Example: 4-neighbor system in 2D image
  - Handling image boundaries
    - Fewer neighbors
    - Neighbors wrapped around

Random Field

- **Clique c** =
  - Subset of sites in S such that
    - c consists of a single site t or
    - Every pair of sites (i,j) in c is such that
      i and j are neighbors of each other
  - Examples
    - Set of cliques of size 1 = C1 = { i | i $\in$ S }
    - Set of cliques of size 2 = C2 = { (i,j) | i $\in$ Nj and j $\in$ Ni }
    - Set of cliques of size 3 = C3 = { (i,j,k) | i $\in$ S, j $\in$ S, k $\in$ S,
      i, j, k are all neighbors of each other }
    - For a 2D image with 4-neighbor system, C3, C4, $\cdots$ are empty
    - For a 2D image with 8-neighbor system, C5, C6, $\cdots$ are empty

Markov Random Field

- **MRF** =
  - A random field with sites S, neighborhood system N, s.t.

    1) **Markovianity**: $P (X_i \mid x_{S-\{i\}}) := P (X_i \mid x_{Ni})$

    - Voxel value is conditionally independent of values at non-neighboring voxels given values at neighboring voxels
    - Does this mean that X_i , X_j independent if i,j not neighbors ?

    2) **Positivity**: $P (x) > 0, \forall x$

    - Positivity ensures that joint PDF/PMF is unique given all local conditional PDFs/PMFs

Andrei Markov

- Mathematician
  - Number theory,
    differential equations,
    probability theory
- Name “Markov” to chains / fields given much after his death (1922)
- PhD Advisor: Chebyshev
- Among best chess players in St. Petersburg
  - Often competed by correspondence

Markov Random Field

- **Homogeneous** MRF =
  - A MRF a homogeneous if
    conditional PDF P (X_i | X_{Ni}) independent of location i
- Interpretation
  - MRF allows us to specify complex (high-D) joint PDF
    via specifying
    simpler (lower-dimensional) conditional PDFs
  - What is the joint PDF ?

Markov-Gibbs Equivalence

- What is the joint PDF ?
  - Hammersely and Clifford (1971) proved that
    X is a MRF on sites S w.r.t. neighborhood system N
    if and only if
    X is a **Gibbs Random Field** on S w.r.t. N

Gibbs Random Field

- GRF =
  - A random field where joint PDF = Gibbs distribution

$$P(x) := \frac{1}{Z} \exp\left(-\frac{1}{T} U(x)\right)$$

where

– Z is a normalization constant called the **partition function**

– T is a constant called the **temperature**

– $U(x)$ is energy function such that

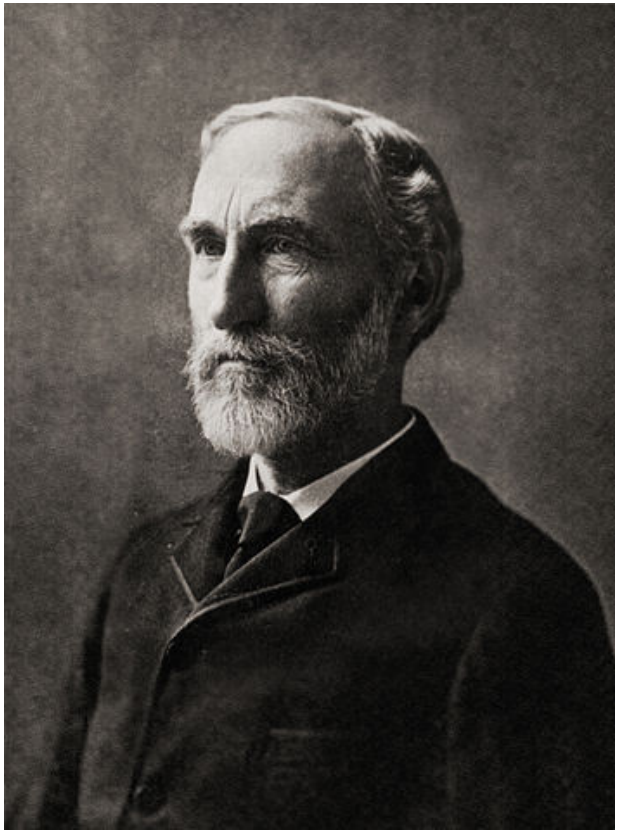
$$U(x) := \sum_{c \in C} V_c(x_c)$$

– where c is a clique in the set of all cliques C

– x_c is the set of image values at sites in the clique c

Josiah Gibbs

- Physics, chemistry, mathematics
  - Thermodynamics
  - Vector calculus
  - Invented the field of statistical mechanics
    - With Maxwell, Boltzmann
- First PhD in engineering
  - Yale University, 1863



Gibbs Random Field

$$P(x) := \frac{1}{Z} \exp\left(-\frac{1}{T}U(x)\right)$$

- $V_c(x_c)$ is called the clique-potential function
- Interpretation
  - Clique-potential functions $V_c(.)$ designed to adapt to tasks
  - Partition function Z difficult to evaluate

$$Z := \sum_x \exp\left(-\frac{1}{T}U(x)\right)$$

- Homogeneous GRF
  - $V_c(x_c)$ independent of location of clique c
- Isotropic GRF
  - $V_c(x_c)$ independent of spatial orientation of clique c

Gibbs Random Field

$$P(x) := \frac{1}{Z} \exp\left(-\frac{1}{T} U(x)\right)$$

- Temperature T controls the sharpness of the distribution
  - T = ∞ → every image x has equal probability P(x), i.e., uniform distribution
  - T = 0 → non-zero probability to those images that were most probable at non-zero temperatures
  - In our applications, by default, T = 1

Image Priors

$$P(x) := \frac{1}{Z} \exp\left(-\frac{1}{T} U(x)\right)$$

- Temperature T
  - Low T → Global maximum x* dissimilar from other x
  - High T → Global maximum x* similar to other x

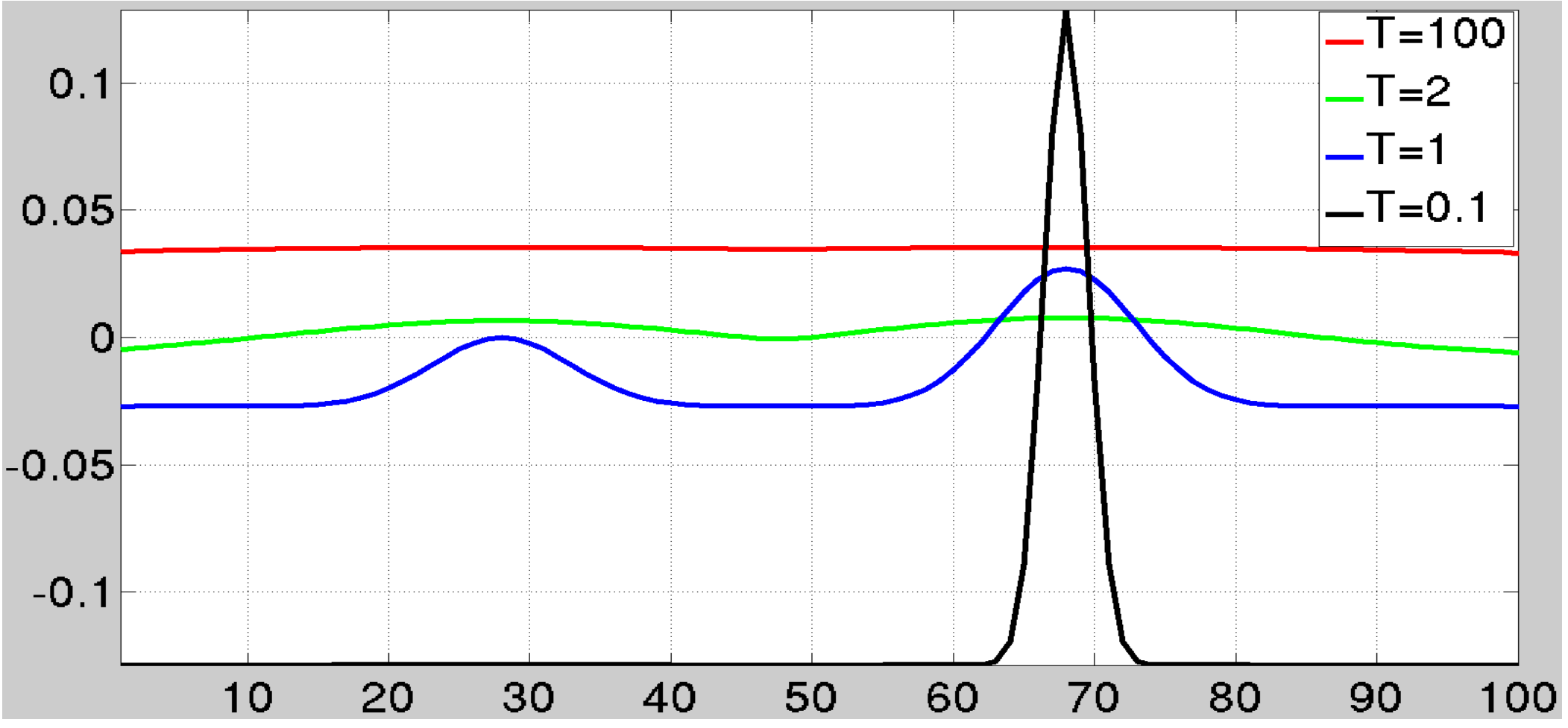
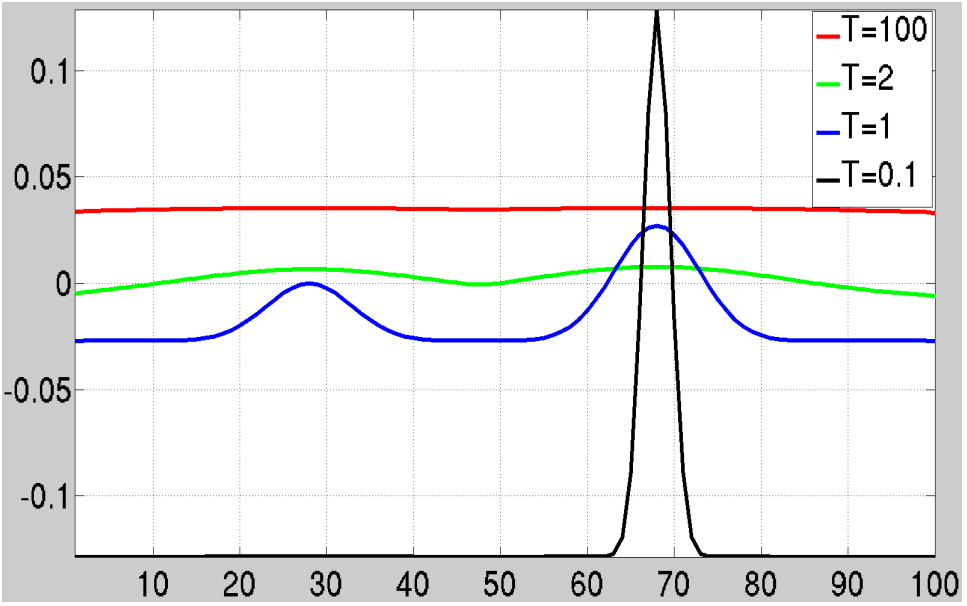


Image Priors

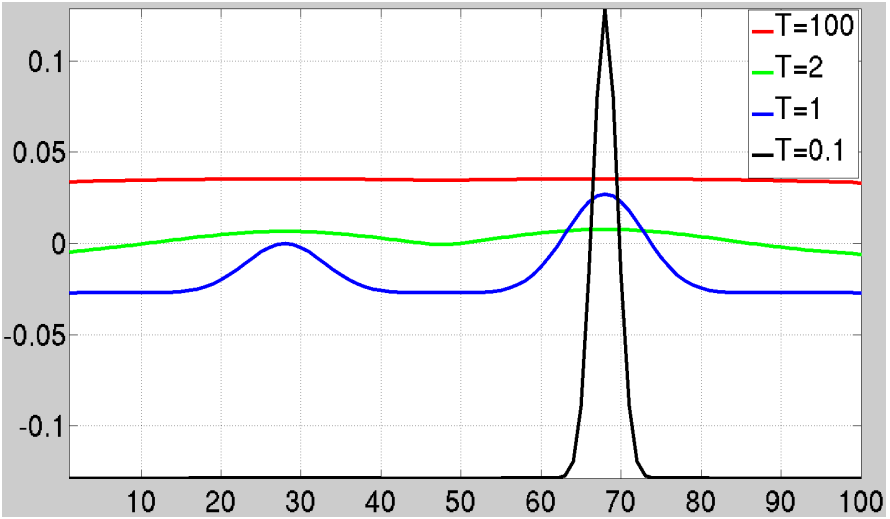
- Simulated annealing
  - Randomized algorithm for optimization
    - Focused on nonlinear optimization problems
  - Inventors
    - [Kirkpatrick et al. 1983 Science]
    - [Cerny 1985 Journal of Optimization Theory & Applications]

Image Priors



- **Simulated Annealing**
  - Find max of blue curve
    - Initial solution x
    - Initial temperature T (large)
    - At current T, repeat, N times, following steps:
      - Generate a random trial solution y
        - Sample from (e.g.) isotropic/symmetric PDF P(Y) around current x (e.g., Gaussian)
      - If P(y) >= P(x)
        - Update solution x → y
      - If P(y) < P(x)
        - Update solution x → y with probability P(y) / P(x)
    - If T < 0.1 (small positive number < 1)
      - Then: Stop. Return x as solution.
    - Else
      - Reduce T. Repeat.

Image Priors



- Simulated Annealing
  - For T large
    - All x have similar probabilities
    - **Initial solution shouldn't matter much**
      - Unlike gradient ascent
  - For T medium
    - Can still move from high-prob state → low-prob state → high-prob state
    - **Helps prevent getting stuck in local maximum**
      - Unlike gradient ascent
  - At T << 1
    - Probability distribution is concentrated at global maxima
    - **Final solution should be one of those**

Image Priors

- Simulated Annealing
  - Why **not** start with T = 0 ?
    - If x isn't at global maximum, then can get lost
  - Why named "**annealing**" ?
    - In metallurgy, process of **slowly cooling** a material
    - Wikipedia: "Annealing can induce ductility, soften material, *relieve internal stresses, refine the structure* by making it homogeneous, and improve cold working properties."
    - Wikipedia: "Annealing glass is critical to its durability. If glass is not annealed, it will retain many of the thermal stresses caused by quenching and significantly decrease the overall strength of the glass."

Image Priors

- Simulated Annealing
  - Algorithm for reducing temperature T ?
    - “Cooling schedule”
    - T(n) non-increasing function over iteration number ‘n’
    - Good performance (convergence to global maxima) requires reducing T very slowly
      - Time consuming
      - T(n) = d / log (n)
    - Very time consuming

Image Priors

- Simulated Annealing
  - In practice
    - Behavior is problem dependent
      - Outperformed best known heuristics for some problems
    - Used a lot in image processing in 1990s for optimization
      - Similar stochastic algorithms used today for sampling-driven statistical inference
        - e.g., machine learning, neural networks
    - **Theoretical guarantee** of finding global maximum comes at a huge **computational cost**
      - Trade off
      - Not useful if slower than exhaustive search

Image Priors

- Simulated Annealing
  - Demo:

    http://upload.wikimedia.org/wikipedia/commons/d/d5/Hill_Climbing_with_Simulated_Annealing.gif

Local Conditionals from GRF

- For a GRF,
  how to get conditional PDF P (X_i | X_{Ni}) ?
  - Given joint PDF: $P(x) := \frac{1}{Z} \exp\left(-\sum_{c \in C} V_c(x_c)\right)$
  - Strategy:
    - Divide set of cliques C into
      2 mutually-exclusive and exhaustive sets of cliques A_i and $C - A_i$, where

      A_i = set of all cliques a containing site i
  - After some rewriting,

$$P(X_i | X_{S-\{i\}}) = \frac{\exp\left(-\sum_{a \in A_i} V_a(X_a)\right)}{\sum_{x_i'} \exp\left(-\sum_{a \in A_i} V_a(X_a)\right)} = \frac{\exp\left(-\sum_{a \in A_i} V_a(X_a)\right)}{Z_i}$$

Local Conditionals from GRF

- For a GRF,
  how to get conditional PDF P (X_i | X_{Ni}) ?

$$P(X_i|X_{S-\{i\}}) = \frac{\exp\left(-\sum_{a\in A_i} V_a(X_a)\right)}{\sum_{x_i'} \exp\left(-\sum_{a\in A_i} V_a(X_a)\right)} = \frac{\exp\left(-\sum_{a\in A_i} V_a(X_a)\right)}{Z_i}$$

  - Interpretation
    - P (X_i | $X_{S-\{i\}}$) only depends on cliques that contain i,
      i.e., sites j that are neighbors of i
    - Denominator = normalization constant ; NOT a function of x_i
    - This also proves that every GRF is a MRF

Example

- **Image smoothing**
  - GRF = MRF with squared-difference potential function
  - **For simplicity, ignore noise model (likelihood)**
  - Smoothing strategy
    - Update image intensity at pixel 'p' ,
      given intensities at neighbors of 'p',
      based on gradient descent
    - Equate derivative of log-prior to zero … solve …
    - What happens at object boundaries ?
      - Heavy smoothing
        - Will happen (to an undesirable extent) even if likelihood was introduced
      - How to avoid this ?

GRF: Adaptive

- **Discontinuity** =
  - is likely when differences between neighboring pixel values is large. e.g., at boundaries of objects in images
- **Outlier** =
  - is a data point that is far from the 'cluster'
  - May be due to noise
- Both scenarios exhibit a large deviation of some kind

GRF: Discontinuity Adaptive

- We want a GRF model that is:

  1) Adaptive to **discontinuities**

  - e.g., if we don’t average pixel intensities belonging to different objects, we blur less

  2) Robust to **outliers**

  - e.g., if we ignore (or weigh down) the outlier, then we are affected less

GRF: Discontinuity Adaptive

- Motivation
  - Consider problem of estimating 'x'
    when observed data in neighborhood of x is

    $y_i = x + \eta_i$ for $i = 1, \cdots, N$ where

    η_i = deviation of neighbor's intensity
    due to **discontinuity or noise**
  - **Assumptions**

    1) Large η_i occur more often due to discontinuity (not noise)

    2) Noise level is smaller than edge strengths

GRF: Discontinuity Adaptive

- Motivation
  - Consider estimating x by minimizing
    sum of penalties of deviation between x and data y_i

$$E(x) := \sum_i g(y_i - x) = \sum_i g(\eta_i(x)) \text{ where } \eta_i(x) := y_i - x$$

  - **Assumption**
    - **g(u) is an even function**
      - g(u) is real valued and g(u) = g(−u)
      - Image: Treat / penalize positive deviations edges same as negative edges
      - Noise: zero mean, symmetric, i.i.d.

GRF: Discontinuity Adaptive

- Motivation
  - When g(u) is even,
    g(u) can be written as a function of $|u|^2$
  - Let **g (u) := H ($|u|^2$)**
  - Plan to optimize based on gradients
  - By chain rule (when u is real-valued):

$$\frac{\partial g(u)}{\partial u} = \frac{\partial H(|u|^2)}{\partial u} = \frac{\partial H(|u|^2)}{\partial (|u|^2)} \frac{\partial |u|^2}{\partial u} = 2uh(u) \text{ where } h(u) := \frac{\partial H(|u|^2)}{\partial (|u|^2)}$$

GRF: Discontinuity Adaptive

- Motivation
  - **Optimization Strategy 1**
    - Compute the gradient of E(x) and equate it to zero
    - Optimal x is :

$$x = \frac{\sum_i h(\eta_i) y_i}{\sum_i h(\eta_i)}$$

  - Insight 1
    - h(·) acts as a weighting function, or an interaction function

$$E(x) := \sum_i g(y_i - x) = \sum_i g(\eta_i(x)) \text{ where } \eta_i(x) := y_i - x$$

GRF: Discontinuity Adaptive

- Motivation
  - **Optimization Strategy 2**
    - Gradient-descent update on x with a specific step-size τ
    - Updated x $= x - \tau \sum_i 2(x - y_i)h(\eta_i)$
  - Insight 2
    - Amount of smoothing $\propto$ 2 (x − yi) h(ηi)

$$E(x) := \sum_i g(y_i - x) = \sum_i g(\eta_i(x)) \text{ where } \eta_i(x) := y_i - x$$

GRF: Discontinuity Adaptive

- Motivation
  - Quadratic penalty : $g(u) := H(|u|^2) = |u|^2$
  - Then, $h(u) = 1$
  - As deviation $|u| \rightarrow \infty$ (or becomes very large)

    (1) weight h(u) remains non-zero

    (2) amount of smoothing $\propto$ 2uh(u), remains infinite
  - So, quadratic penalty blurs discontinuities heavily
  - How to design g(u) or h(u) (interaction function) to denoise while preserving discontinuities ?

GRF: Discontinuity Adaptive

- Rules for Designing the Interaction Function h(u)
  - (0) Real valued
    - It is derived from a penalty g(u) that is real valued
  - (1) Continuous
  - (2) Non negative
    - Weighting function, want convex updates
  - (3) Even function
    - Weighted sum of a positive deviation and a negative deviation should cancel out
  - (4) Non-increasing
    - Want larger deviations u to produce not-greater weights h(u)

GRF: Discontinuity Adaptive

- Rules for Designing the Interaction Function h(u)

(5.1) h(u) should $\rightarrow$ 0, when u $\rightarrow \infty$

- Want the interaction / weight = 0 when for infinite deviation

(5.2) $\lim_{u \rightarrow \infty} g'(u) := \lim_{u \rightarrow \infty} 2\, u\, h(u) \leq C$,
where $0 \leq C < \infty$ is a constant,

- 5.2 is a stronger version of 5.1 (5.2 implies 5.1)
- As deviation u $\rightarrow \infty$, we want penalty g(u) to either :

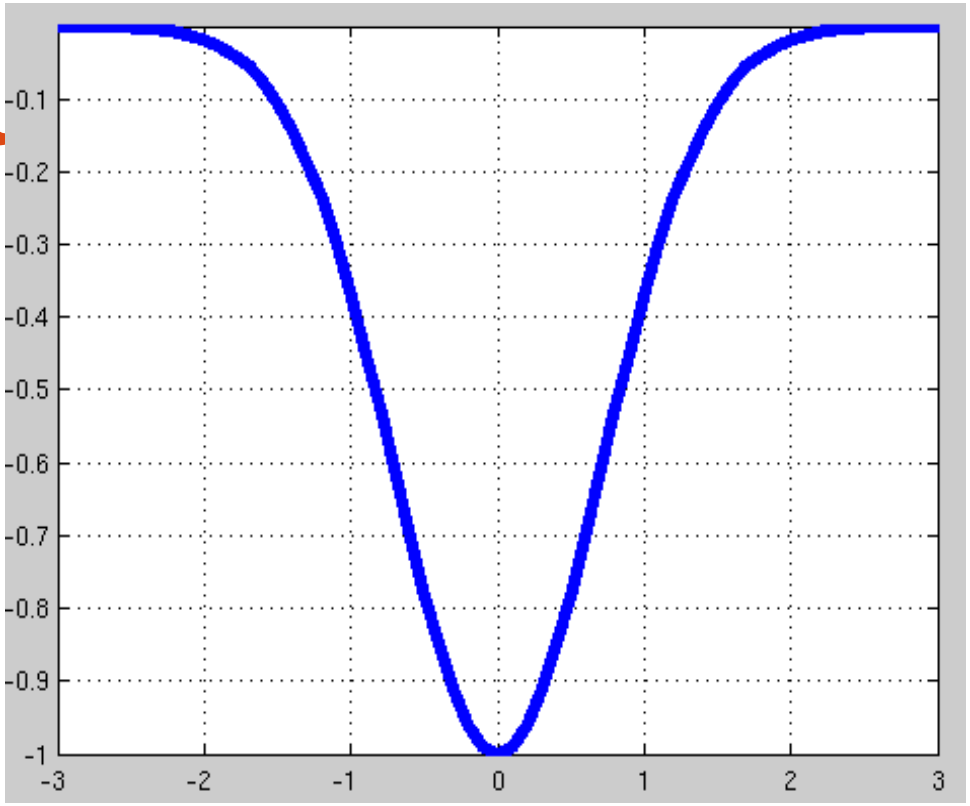
(i) C = 0 case :
Penalty g(u) is constant / bounded.
Zero interaction = NO smoothing (weight zero, h(∞) = 0).

(ii) C > 0 case :
Penalty g(u) increases at sub-linear rate.
Bounded smoothing. (amount of smoothing $\propto$ 2 u h(u) ≤ C)

GRF: Discontinuity Adaptive

- Example Penalty Function (strictly convex, for real u)
  - Quadratic: $g(u) := |u|^2$
  - What is h(u) ?
    - $g'(u) = 2u$. So, $h(u) = 1$
  - h(u) satisfies Conditions 1, 2, 3, 4
  - h(u) violates Condition 5.1 and 5.2
    - g(u) doesn't respect discontinuities ; causes excessive smoothing / blurring

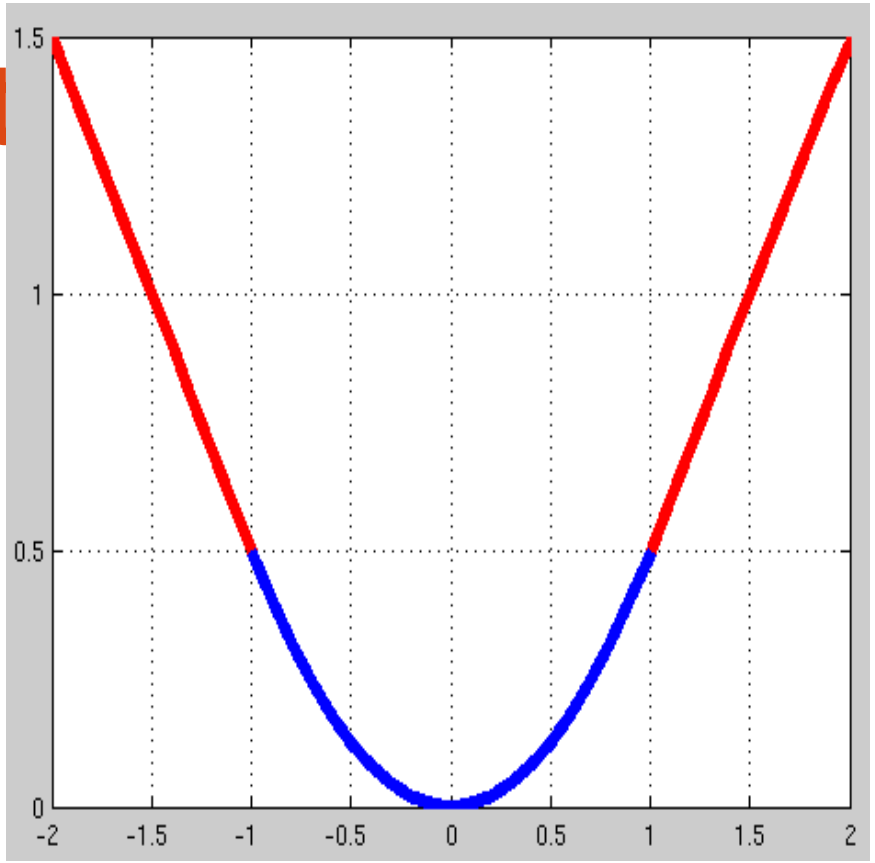
GRF: Discontinu



- Example Penalty Function (non convex, for real u)
  - $g(u) := -\gamma \exp(-|u|^2/\gamma)$

    0 < gamma < ∞ is user-defined constant
  - What is h(u) ?
  - $\frac{\partial g(u)}{\partial u} = 2u \exp(-|u|^2/\gamma)$ and $h(u) = \exp(-|u|^2/\gamma)$
  - h(u) satisfies Conditions 1, 2, 3, 4
  - h(u) satisfies Condition 5.1 : $\lim_{u\to\infty} \exp(-|u|^2/\gamma) = 0$
  - h(u) satisfies Condition 5.2 : 0 = $\lim_{u\to\infty} u \exp(-|u|^2/\gamma)$

GRF: Discontinuit



- Example Penalty Function (convex, for real u)
  - Huber Function:

$$g(u) := \frac{1}{2}|u|^2 \text{ when } |u| \leq \gamma$$

$$g(u) := \gamma|u| - \frac{\gamma^2}{2} \text{ when } |u| > \gamma$$

  - What is h(u) ?

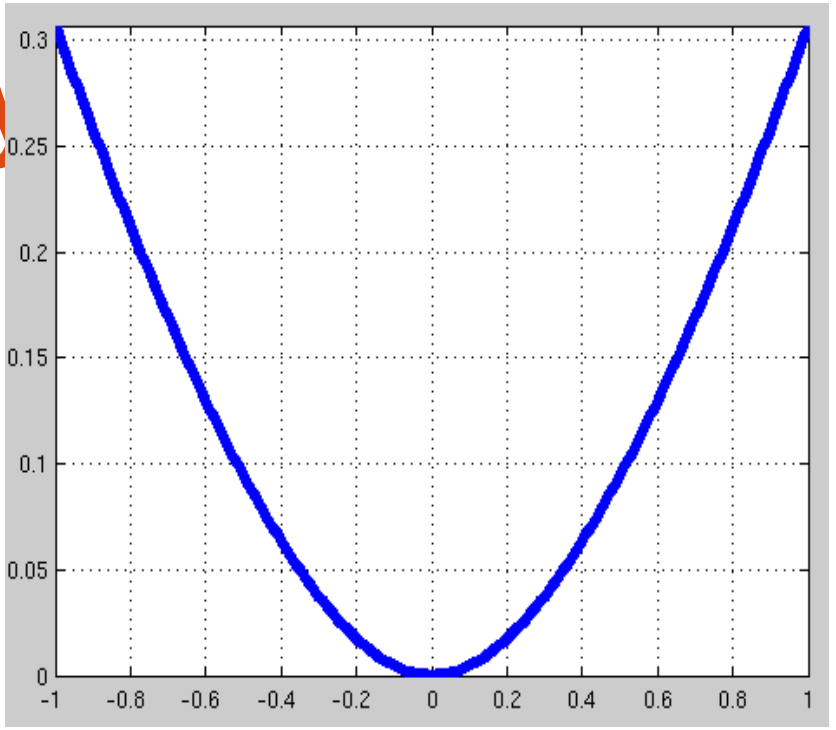
$$\frac{\partial g(u)}{\partial u} = u \text{ when } |u| \leq \gamma \qquad h(u) = \frac{1}{2} \text{ when } |u| \leq \gamma$$

$$\frac{\partial g(u)}{\partial u} = \gamma \operatorname{sgn}(u) \text{ when } |u| > \gamma \qquad h(u) = \frac{\gamma}{2|u|} \text{ when } |u| > \gamma$$

  - h(u) satisfies Conditions 1, 2, 3, 4, 5.1, 5.2

GRF: Discontinuity

- Example Penalty Function
  (strictly convex, for real u)
  - $g(u) := \gamma|u| - \gamma^2 \log\left(1 + \frac{|u|}{\gamma}\right)$
  - What is h(u) ? $h(u) = \frac{\gamma}{2(\gamma + |u|)}$
  - h(u) satisfies Conditions 1, 2, 3, 4, 5.1, 5.2



Bernoulli family

- Family of merchants and scholars
- Over 3 generations, 8 Bernoullis contributed to foundations of applied mathematics and physics

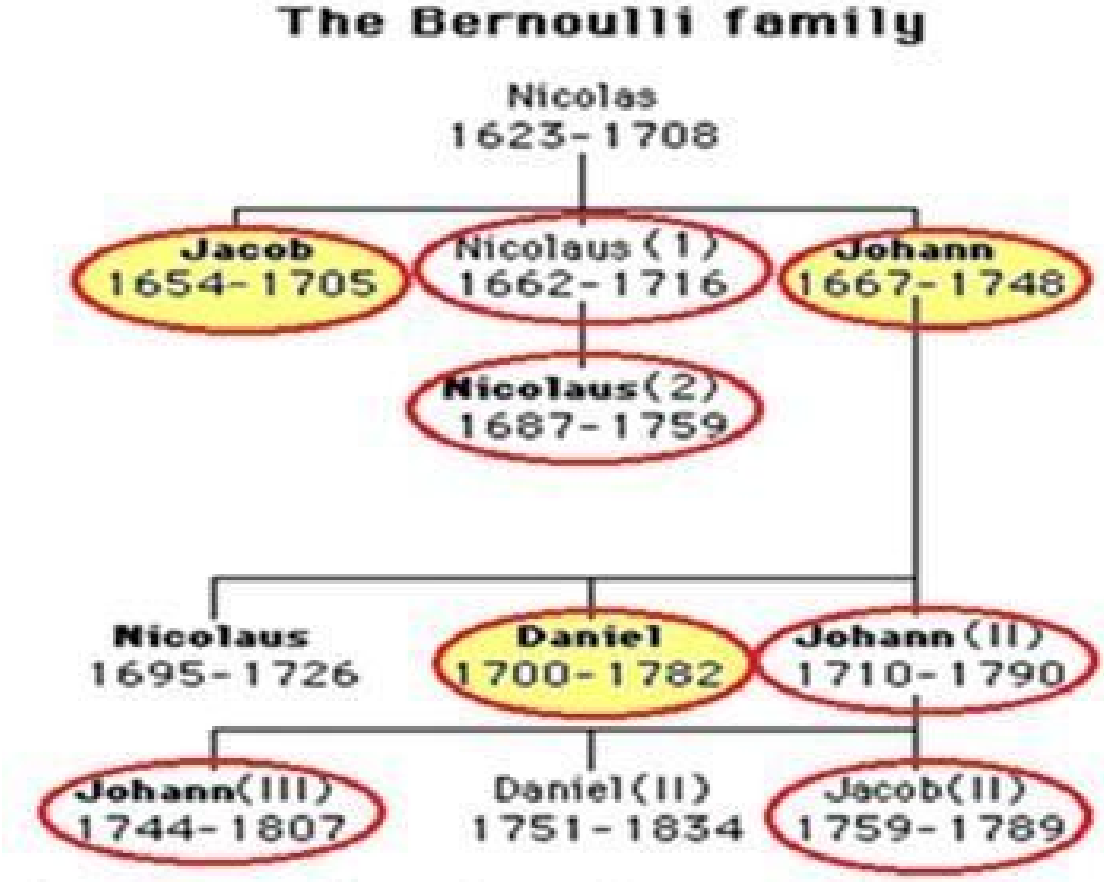


Fig 1. The Bernoulli Family Tree (the ones circled in red are mathematician members)