Priors on Images

Suyash P. Awate

Bayes Rule

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

Bayes Rule

- Bayes Rule
  - Application : Denoising
    - P (NoiselessImage | NoisyImage)
      = P (NoisyImage | NoiselessImage)
      * **P (NoiselessImage)**
      / P (NoisyImage)
    - Likelihood PDF = P (NoisyImage | NoiselessImage)
      - Noise model
    - Prior PDF = **P (NoiselessImage)**
      - **Our prior beliefs about the noiseless image**
    - Posterior PDF = P (NoiselessImage | NoisyImage)
      - Product of likelihood and prior (with normalization factor)

Bayes Rule

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

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 and 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 : $\mu = \frac{\sigma_0^2 \bar{u} + \sigma^2 \mu_0 / N}{\sigma_0^2 + \sigma^2 / N}$
  - 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)

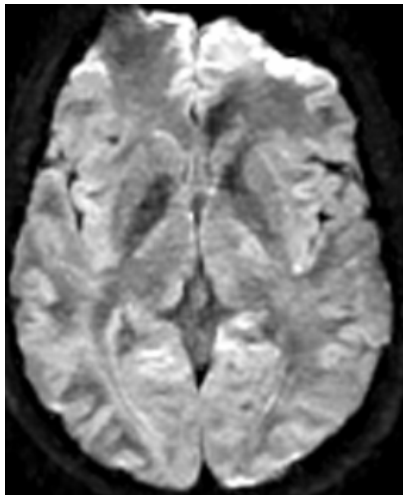
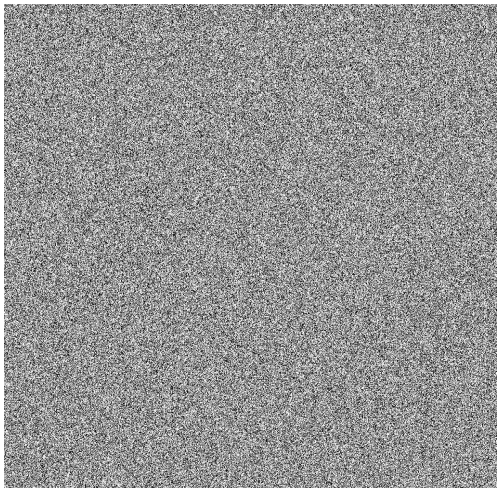
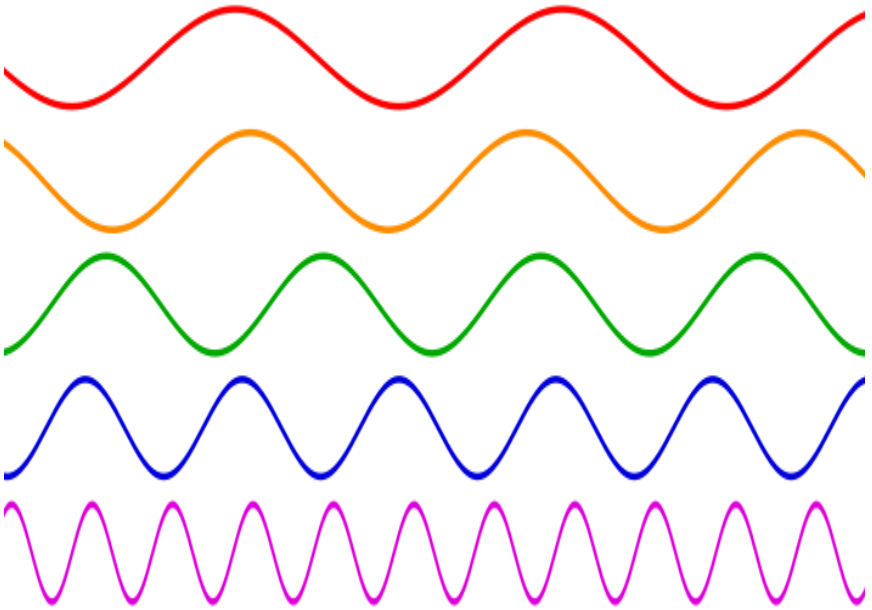
      What if $\sigma_0 \to 0$? (strong prior: ignore the data)

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
  - Image values are spatially (piecewise) smooth
  - Discontinuities / large changes possible only at object boundaries

Image Prior

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

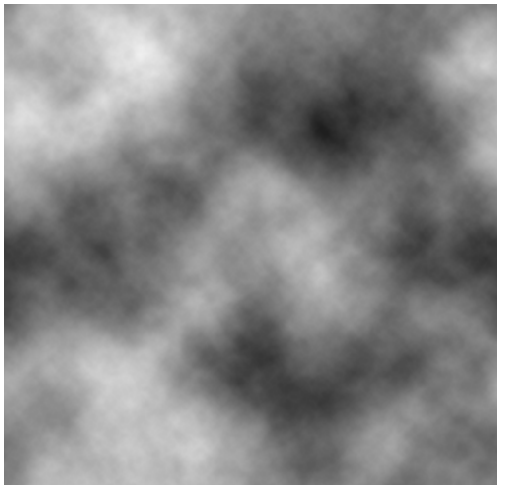


Image Prior

- Assign a probability of occurrence to every image within the image space
  - What if data = image in row 1 col 3 ?
  - What if data = image in row 2 col 1 ?

Image Prior

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

Random Field

- Random field =
  - Given a probability space (Ω, F, P),
    a random field is a collection of random variables { Xi }
    indexed by values i in a topological space S
- 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
- Observed image x = single realization of the random field X under some PDF P(X)

Random Field

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

    where N_i is the set of sites neighboring site i
- Neighboring relation =

  (1) A site isn't a neighbor to itself :
  $i \notin N_i$

  (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 are neighbors of each other
  - Examples
    - Set of cliques of size 1 = C1 = { i | i ∈ S }
    - Set of cliques of size 2 = C2 = { (i, j) | i ∈ Ni and j ∈ Nj }
    - Set of cliques of size 3 = C3 = { (i, j, k) | i ∈ S, j ∈ S, k ∈ S, i, j, k are all neighbors of each other }
    - For a 2D image with 4-neighbor system, C3 , C4 , · · · are all empty !!
    - For a 2D image with 8-neighbor system, C5 , C6 , · · · are all empty !!

Markov Random Field

- MRF =
  - A random field with sites S, neighborhood system N

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

    (2) Markovianity : $P(Xi \mid x_{S-\{i\}}) := P(Xi \mid X_{Ni})$
  - Voxel values are conditionally independent of values at all non-neighboring voxels when the values at the neighboring voxels are given
  - Positivity ensures that the joint PDF/PMF is unique given all the 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 (Xi | X_{Ni}) independent of location i
- Interpretation
  - MRF allows us to specify the complex (high-dimensional) 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

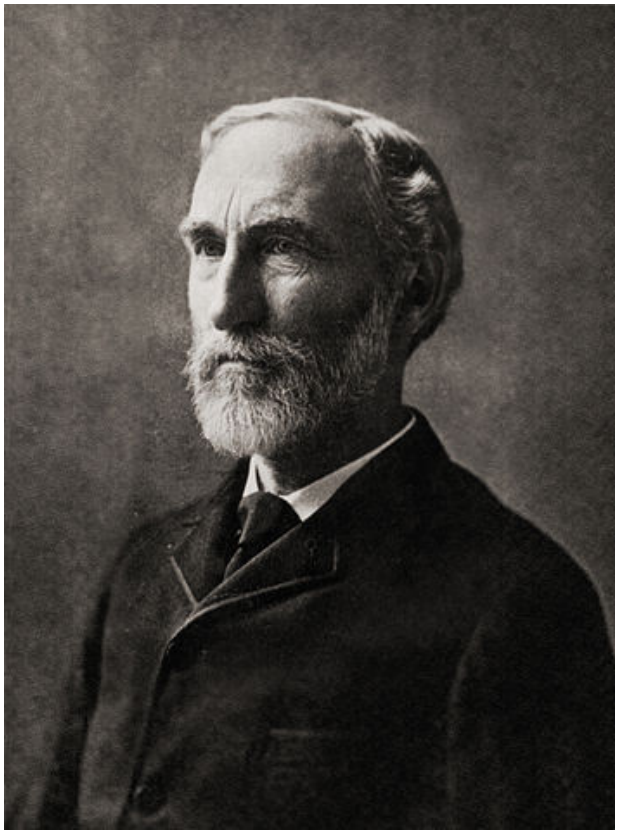
- GRF =
  - A random field where the joint PDF = Gibbs distribution
  - $$P(x) := \frac{1}{Z} \exp\left(-\frac{1}{T} U(x)\right)$$

    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 doctorate in engineering
  - Yale University, 1863

Gibbs Random Field

- Vc (xc) is called the clique-potential function
- Interpretation
  - No restrictions on the clique-potential function Vc (.)
  - Partition function Z difficult to evaluate

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

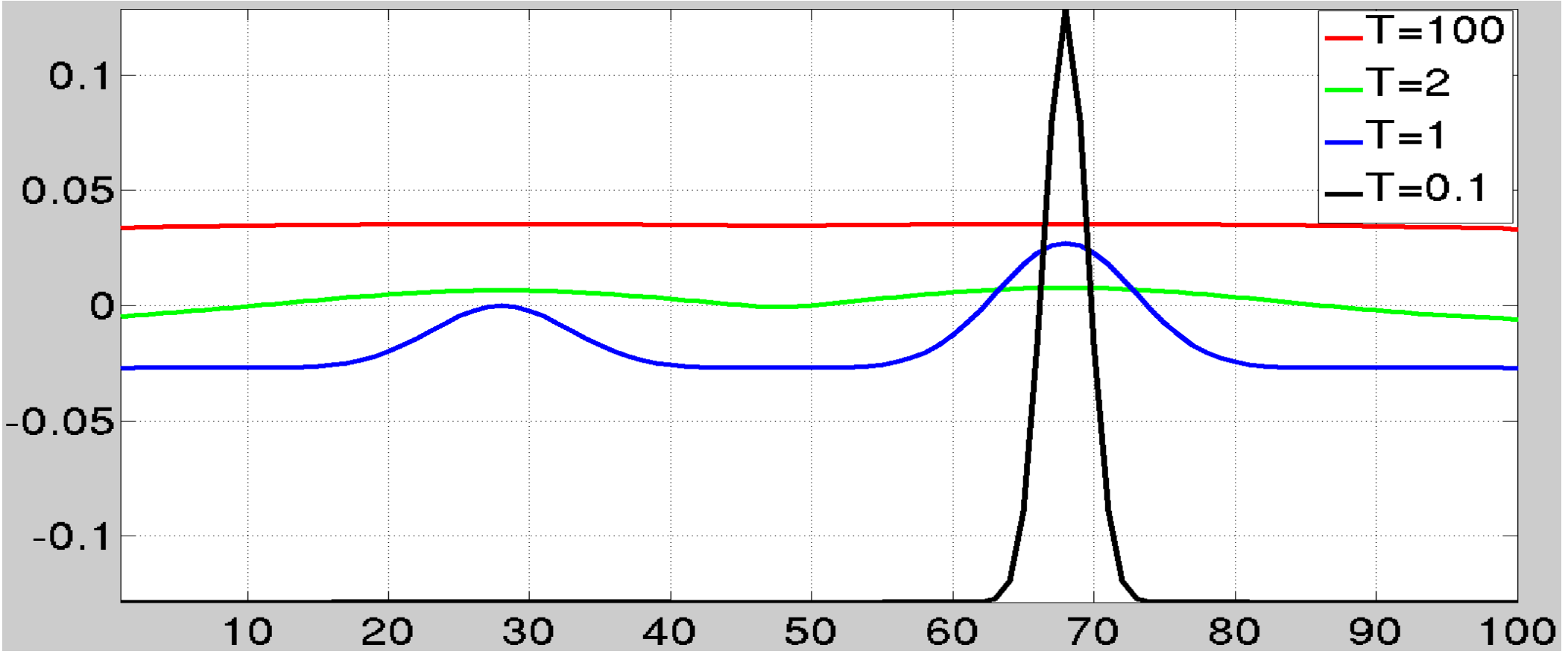
- Homogeneous GRF
  - Vc (xc) independent of location of clique c
- Isotropic GRF
  - Vc (xc) independent of spatial orientation of clique c

Gibbs Random Field

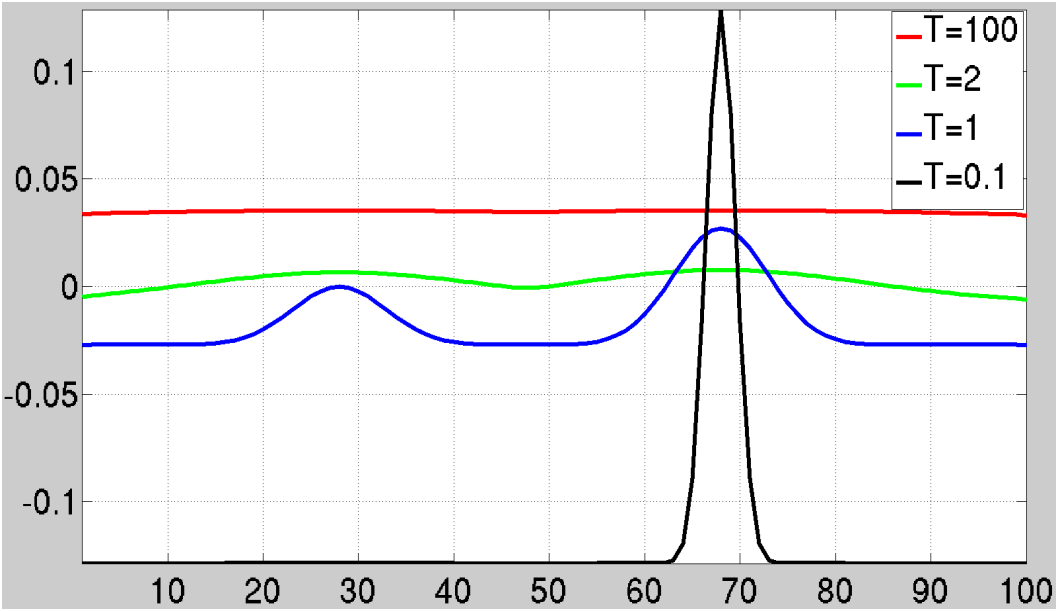
- Temperature T controls the sharpness of the distribution
  - Infinite T gives every image x the same probability P(x), i.e., uniform distribution
  - Zero T gives non-zero probability to the images that were most probable at non-zero temperatures
  - In our applications, by default, T = 1

Image Priors

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

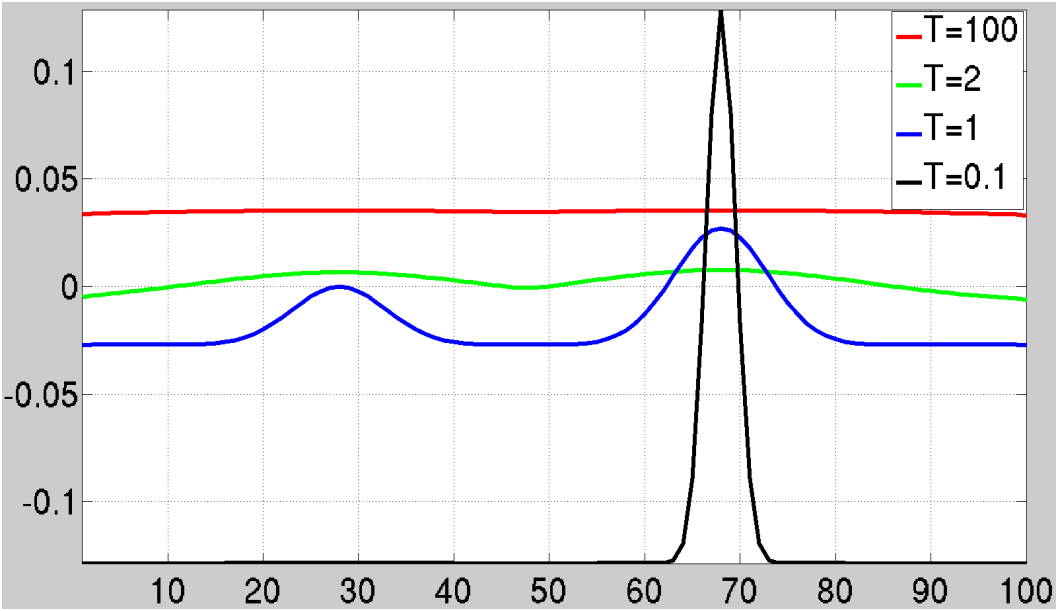


Image



- Simulated Annealing
  - Algorithm:
    - Initial temperature T (large) and initial solution x*
    - Generate a random trial solution y
      - Sample from uniform distribution over all x
    - If P(y) >= P(x*)
      - Update solution x* to y
    - If P(y) < P(x*)
      - Update solution x* to y with probability P(y) / P(x*)
    - If T = 0, then stop. Return x* as solution.
    - Reduce T and repeat

Image



- Simulated Annealing
  - For T large
    - All x have similar probabilities
    - **Initial solution shouldn't matter much !!**
  - For T medium
    - Can move from high-prob state to low-prob state
    - **Helps to prevent getting stuck in local maximum !!**
  - At T=0
    - 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 stuck
  - 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”
    - Non-increasing function over iteration number
    - 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
    - Was used extensively in image processing in 1980s and 90s
    - **Theoretical guarantee** of finding global maximum comes at a huge **computational cost**
      - Trade off
      - Not useful if not faster than exhaustive search !
    - Inventors
      - [Kirkpatrick et al. 1983 Science]
      - [Cerny 1985 Journal of Optimization Theory & Applications]

Image Priors

- Simulated Annealing
  - Demo

    http://en.wikipedia.org/wiki/File:Hill_Climbing_with_Simulated_Annealing.gif

Local Conditionals from GRF

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

      Ai = the 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 the conditional PDF P (Xi | 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 (Xi |$X_{S-\{i\}}$) only depends on the cliques that contain i, i.e., the sites j that are neighbors of i
    - Denominator is the normalization constant that is NOT a function of xi
    - 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
      - How to avoid this ?

GRF: Adaptive

- Discontinuity =
  - is likely when the 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
  - Adaptive to discontinuities
    - e.g., if we don’t average pixel intensities belonging to different objects, we blur less
  - Robust to outliers
    - e.g., if we ignore (or weigh down) the outlier, then we are affected less

GRF: Discontinuity Adaptive

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

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

    η_i = deviation of the neighbor's intensity due to discontinuity or noise
  - Assumption
    - Large η_i occur more often due to discontinuity (not noise)
    - Noise level is smaller than edge strengths

GRF: Discontinuity Adaptive

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

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

  - Assumption
    - g(u) is an even function
      - g(u) is real valued and g(u) = g(−u)
      - Image: Treat positive deviations edges the same as negative edges
      - Noise: zero mean, 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 :

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

GRF: Discontinuity Adaptive

- Motivation
  - Optimization Strategy 1
    - Compute the gradient of this function 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 !!

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 ∝ 2 (x − yi) h(ηi)

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 $\propto 2u$, remains infinite
  - Thus, 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) $\lim_{u \rightarrow \infty} g'(u) := \lim_{u \rightarrow \infty} 2\,u\,h(u) = C$,

  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 :

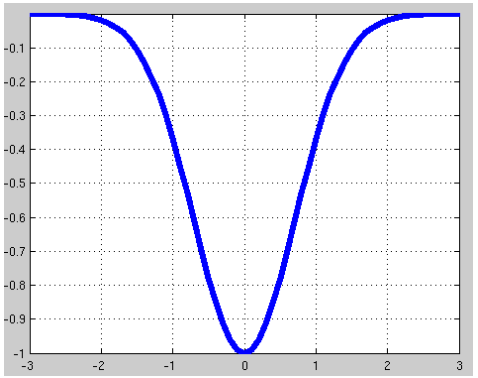
    Penalty g(u) increases at sub-linear rate.

    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

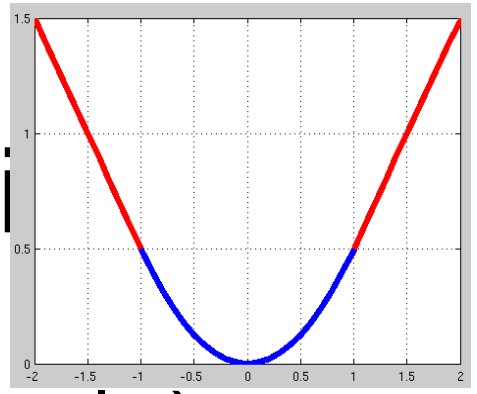
GRF: Discontinuity Adapt



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

    0 < gamma < ∞ is user-defined positive finite 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: Discontinuity Adapti



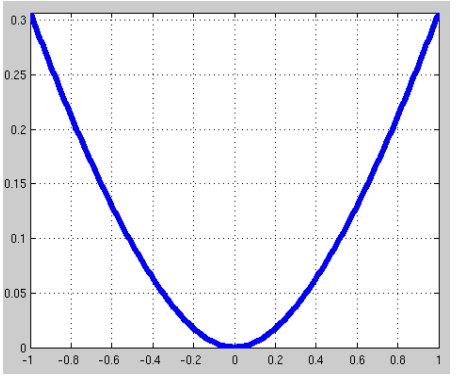
- Example Penalty Function (convex, for real u)
  - Huber Function : $g(u) := \frac{1}{2}|u|^2$ when $|u| \leq \gamma$

    $g(u) := \gamma|u| - \frac{\gamma^2}{2}$ when $|u| > \gamma$
  - What is h(u) ?

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

    $\frac{\partial g(u)}{\partial u} = \gamma\,\mathrm{sgn}(u)$ when $|u| > \gamma$ $\qquad h(u) = \frac{\gamma}{2|u|}$ when $|u| > \gamma$
  - h(u) satisfies Conditions 1, 2, 3, 4, 5.1, 5.2

GRF: Discontinuity Adapti



- 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