Political Methodology Summer Meeting
July 2004, Stanford University

# Dynamic Tempered Transitions for Exploring Multimodal Posterior Distributions

**JEFF GILL**
Department of Political Science
University of California, Davis

**GEORGE CASELLA**
Department of Statistics
University of Florida

# Project Description

- It is well-known that highly-multimodal target distributions are problematic for basic MCMC algorithms (and exacerbated by high dimensions too).

- Of course they are also highly problematic for standard maximum likelihood numerical algorithms: quasi-Newton method BFGS, standard and modified Newton-Raphson, steepest descent, IWLS, etc.

- Key problem: algorithms are attracted to isolated local maxima and either take a long time to leave, or in some cases never leave during the time the chain path is recorded.

- Our purpose: to provide a new algorithm that efficiently explores multimodal posterior distributions.

# The Metropolis-Hastings Algorithm

- A type of stochastic process that will help us describe posterior distributions empirically.

- Background:
  - The original work by Metropolis et al. postulated a two-dimensional enclosure with $n = 10$ molecular particles.
  - They sought to estimate the state-dependent total energy of the system at equilibrium.
  - Of course there is an incredibly large number of locations for the molecules in the system that must be accounted for and this number grows exponentially with time.
  - Their idea is to *simulate* this system probabilistically by generating moves that are more likely than others based on positions that are calculated using uniform probability generated candidate jump points.
  - Therefore the simulation produces an estimated force based on a statistical, rather than deterministic, arrangement of particles.

# Metropolis Circa 1953

- Assumptions:
  - We want to describe a posterior, $\pi(\boldsymbol{\theta})$, which is difficult to do analytically.
  - Candidate values will be generated from the distribution $q(\boldsymbol{\theta}'|\boldsymbol{\theta})$ where $q(.|\boldsymbol{\theta})$ a valid (and convenient) PDF for admissible values of $\boldsymbol{\theta}$.
  - Also assume for now that this candidate generating (instrumental, jumping, or proposal) distribution is symmetric in its arguments:

    $$q(\boldsymbol{\theta}|\boldsymbol{\theta}') = q(\boldsymbol{\theta}'|\boldsymbol{\theta}).$$

    Otherwise the Markov chain is not irreducible (irreducible: "all reasonable sized sets can be reached from every possible starting point" –Meyn and Tweedie 1993).
  - Note that the support of $\pi(\boldsymbol{\theta})$ and $q(\boldsymbol{\theta}'|\boldsymbol{\theta})$ must be equivalent.

# Metropolis Circa 1953 (cont.)

- A a single Metropolis iteration from the symmetric form has the following steps:

  1. Sample $\boldsymbol{\theta}'$ from $q(\boldsymbol{\theta}'|\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is the current location.
  2. Sample $u$ from $\mathcal{U}[0:1]$.
  3. If
     $$a(\boldsymbol{\theta}', \boldsymbol{\theta}) = \frac{\pi(\boldsymbol{\theta}')}{\pi(\boldsymbol{\theta})} > u$$
     then accept $\boldsymbol{\theta}'$.
  4. Otherwise keep $\boldsymbol{\theta}$ as the subsequent draw.

- The result is the chain:
  $$\boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_n$$
  where consecutive values are not necessarily unique.

- *Problem:* it turns out that the symmetric requirement of the instrumental distribution is an annoying restriction.

# Metropolis-Hastings Circa 1970

- Background:
  - Hastings (1970) as well as Peskun (1973) generalized the Metropolis et al. version by suggesting a way to use other jumping distributions.
  - Question: can an asymmetric instrumental distribution work if there is some sort of compensation in the acceptance ratio?
  - Yes (obviously), but there are conditions.

# Metropolis-Hastings Circa 1970

- Generalizing:
  - Define $A(\boldsymbol{\theta}', \boldsymbol{\theta})$ as the *actual transaction function*, which differs from $q(\boldsymbol{\theta}'|\boldsymbol{\theta})$ because it includes the accept/reject decision:

$$A(\boldsymbol{\theta}', \boldsymbol{\theta}) = q(\boldsymbol{\theta}'|\boldsymbol{\theta}) \min\left\{\frac{\pi(\boldsymbol{\theta}')}{\pi(\boldsymbol{\theta})}, 1\right\}$$

  - Now require that the transition kernel satisfy the *detailed balance equation* instead of symmetry:

$$A(\boldsymbol{\theta}', \boldsymbol{\theta})\pi(\boldsymbol{\theta}') = A(\boldsymbol{\theta}, \boldsymbol{\theta}')\pi(\boldsymbol{\theta}),$$

    (also called the *reversibility condition*).
  - Under this condition the actual transaction function becomes:

$$A(\boldsymbol{\theta}', \boldsymbol{\theta}) = q(\boldsymbol{\theta}'|\boldsymbol{\theta}) \min\left\{\frac{q(\boldsymbol{\theta}|\boldsymbol{\theta}')}{q(\boldsymbol{\theta}'|\boldsymbol{\theta})}\frac{\pi(\boldsymbol{\theta}')}{\pi(\boldsymbol{\theta})}, 1\right\}$$

# Metropolis-Hastings Circa 1970

- What does this buy you? The following:

  1. If the marginal distribution of the chain is proper, then this is a **positive chain** *(Diaconis and Stroock 1991, p.36)*.
  2. A positive chain is **positive recurrent** *Aldous and Diaconis 1987, p.70)*.
  3. Reversibility and positive recurrence means that there is a **unique stationary distribution** *(Diaconis, Holmes, and Neal 1998, p.727; Diaconis and Freedman 1980, p.128)*.
  4. **Aperiodicity** and positive recurrence means that the chain will eventually converge to this stationary distribution *(Diaconis and Fill 1990, p.1485-6)*.
  5. **Metropolis** chains are irreducible and aperiodic by construction with u.s.d $\pi(\boldsymbol{\theta})$. *(Diaconis and Saloff-Coste 1995, p.112)*.

# Metropolis-Hastings Circa 1970

- Generalizing, cont.:
  - This means replacing the 1953 acceptance ratio with:

$$a(\boldsymbol{\theta}', \boldsymbol{\theta}) = \frac{\pi(\boldsymbol{\theta}')}{\pi(\boldsymbol{\theta})} > u \qquad \text{with:} \qquad a(\boldsymbol{\theta}', \boldsymbol{\theta}) = \frac{q(\boldsymbol{\theta}|\boldsymbol{\theta}')}{q(\boldsymbol{\theta}'|\boldsymbol{\theta})}\frac{\pi(\boldsymbol{\theta}')}{\pi(\boldsymbol{\theta})} > u.$$

  - So we can use the same Metropolis-Hastings engine but replace symmetry with reversibility.
  - Workhorses:

    **random walk chain**, $\boldsymbol{\theta}' = \boldsymbol{\theta} + f(\tau)$
    (where $f(\tau)$ is some convenient PDF).

    **independence chain**, $\boldsymbol{\theta}' = f(\tau)$
    (ignores current position entirely).

    **hit and run chain**, $\boldsymbol{\theta}' = \boldsymbol{\theta} + Ds \times \mathbf{Dr}$
    (separates direction and distance decisions).

- *Problem:* standard M-H gets can get stuck in isolated modes with low total probability for long periods of time.

# Simulated Annealing

**Basic Idea** (Kirkpatrick, *et al.* 1983):

- Analogous with metallurgy: heat up the MCMC transition matrix such that the chain converges weakly over a near-uniform distribution, then progressively cool.

- Once cooling begins, the chain is observed to converge at progressively declining temperatures until the transition matrix returns to its original state.

- While the Markov chain described here is not *homogeneous* like the standard Metropolis-Hastings algorithm, Hàjek (1988) showed that the discrete version still has convergence properties.

# Simulated Annealing

## Metropolis-Hastings Implementation: For “Heated-Up” Chains:

- Heating the kernel flattens out its probability structure toward a uniform distribution, thus melting down modes.

- As the jumping distribution generates candidate positions, very few of these will be rejected and the Markov chain will rarely stay in place.

- *The Good:* it means that the chain can freely explore the sample space without impediments.

- *The Bad:* there is obviously much less of a tendency to remain in the (previous) high density areas.

- *The Ugly:* there is a cooling schedule trade-off:
  - slow cooling enables greater coverage of the sample space,
  - faster cooling gives more reasonable simulation times.

# Simulated Annealing

**Temperature Schedule:**

- Start with a high initial temperature, $T_0$, that provides sufficient melting.

- Gradually cool down to one by decreasing slightly to $T_t$ at era $t$.

- At each temperature modify the target according to $\pi^*(\boldsymbol{\theta}) = \pi(\boldsymbol{\theta})^{\frac{1}{T}}$.

- Some example cooling schemes:
  - linear: $T_t = T_{t-1} - kt$
  - logarithmic: $T_t = 1 + kT_{t-1}/\log(t+1)$
  - proportional: $T_t = \epsilon T_{t-1}, 0 < \epsilon < 1$
  - geometric: $T_t = 1 + k\epsilon^t T_{t-1}, 0 < \epsilon < 1.$

# Simulated Annealing

## Metropolis Implementation Details:

- During era $t$ we are at temperature $T_t$, starting at time 0 with the heated up temperature $T_0$.

- Subsequent values are chosen according to the algorithm:
  - At the $j^{th}$ step draw $\boldsymbol{\theta}'$ from a convenient distribution around the current position, $\boldsymbol{\theta}^{[j]}$.
  - Define: $a(\boldsymbol{\theta}', \boldsymbol{\theta}) = \exp[(\pi(\boldsymbol{\theta}') - \pi(\boldsymbol{\theta}^{[j]}))/T_t]$, and make the decision:

$$\boldsymbol{\theta}_j^{[t+1]} = \begin{cases} \boldsymbol{\theta}' & \text{with probability} \qquad P\left[\min\left(a(\boldsymbol{\theta}', \boldsymbol{\theta}), 1\right)\right] \\ \boldsymbol{\theta}^{[j]} & \text{with probability} \qquad 1 - P\left[\min\left(a(\boldsymbol{\theta}', \boldsymbol{\theta}), 1\right)\right] \end{cases}$$

- After thermal convergence or sufficient traversal is concluded at this temperature, move down the temperature schedule from $T_t$ to $T_{t+1}$.

- Repeat steps 1-3 until the temperature schedule has been completed.

- *Problem:* only the cold chain is useful for inferential purposes.

# Simulated Annealing

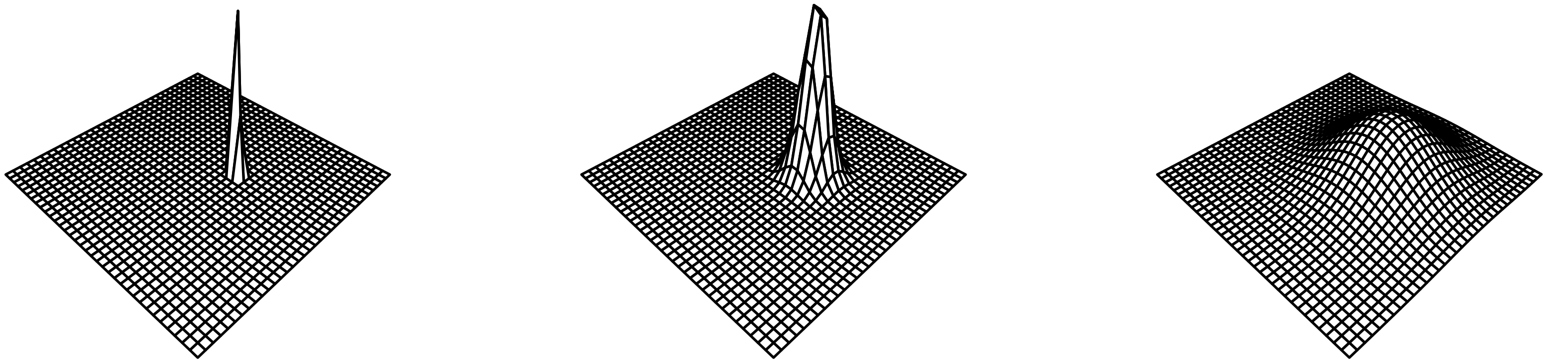
**Example:**

- The infamous *witch's hat distribution* of Matthews (1991):

$$p(\boldsymbol{\theta}|\mathbf{x}) = (1-\delta)[2\pi\sigma^2]^{-d/2}\exp\left[-\sum_{i=1}^{d}\frac{1}{2\sigma^2}(x_i-\boldsymbol{\theta}_i)^2\right] + \delta I_{(0,1)}(x_i),$$

  where $I_{(0,1)}(x_i)$ is an indicator function equal to one when $x_i$ is in the interval $(0,1)$ and zero otherwise.

- Melting Down the Witch's Hat Distribution, at temperatures 1,25,300:



## Alternative 1 Metropolis-Coupled Markov Chain Monte Carlo

- Because of the high-dimensional and multimodal nature of objective functions of interest here, the candidate distribution and the temperature schedule must be chosen with great care to allow adequate exploration of the space.

- One early solution: **MCMCMC** (Geyer 1991). Characteristics (for a vector $\mathbf{c}$ from $\pi(\mathbf{c})$):
  - Run $N$ parallel chains at different heat levels from 1 to $\beta_N$.
  - Thus $N$ transition kernels are defined, $MC_1, \ldots, MC_N$, with stationary distributions $m_1, \ldots m_N$.
  - At time $t$ select two chains, $i$ and $j$, and attempt to swap states:

$$\mathbf{c}_i^{(t)} \Longleftrightarrow \mathbf{c}_j^{(t)}$$

  - with a Metropolis decision using the hotter chain as the baseline, probability: $\min\left\{ \frac{m_i(\mathbf{c}_j^{(t)})/m_i(\mathbf{c}_i^{(t)})}{m_j(\mathbf{c}_j^{(t)})/m_j(\mathbf{c}_i^{(t)})}, 1 \right\}$.
  - Record only the cold chain for inferential purposes.

- *Problem:* $N$ usually needs to be large.

# Alternative 2 Simulated Tempering

- Geyer & Thompson (1995) and Marinari & Parisi (1992) proposed an alternative algorithm called simulated tempering that reduces MCMCMC to a single chain.

- Essentially temperature becomes a discrete random variable (i.e. it now possesses a distributional assignment) so the system can heat *and* cool as time proceeds, at each step draw:

$$[\boldsymbol{\theta}', \beta] \sim q(\boldsymbol{\theta}'|\boldsymbol{\theta})g(\beta)$$

- Why would one do this? Now elderly chains can still avoid being trapped at local maxima without having to run many replicants.

- *Problem:* specifying the marginal distribution for $\beta$ is difficult in practice since $f(\beta)$ needs to favor the cold chain sufficiently to get a large enough sample for inferences, but it also needs to be able to sample from higher temperature values as well for good traversal.

## $\boxed{\text{Alternative 3}}$ Tempered Transitions

- Neal (1996) extends simulated tempering with tempered transitions to heat up the posterior distribution at each step (preserving the detailed balance equation).

- Basic idea: “ladder” up and down at every iteration of the chain with random walk steps.

- Each ladder step specifies a (nonnormalized) stationary distribution defined on the same state space but at progressively hotter temperatures ($\beta_i$) going up.

- We accept the last (bottom) ladder value with a Metropolis decision.

- Details:
  - define a sequence of candidate densities $m_i, i = 1, \ldots, N$, where as $i$ increases the $m_i$ get flatter going up the ladder then again more peaked going down the ladder.
  - parameterize: $m_i = m^{1/\beta_i}$,
  - where: $1 < \beta_1 < \beta_2 < \cdots < \beta_{N-1} < \beta_N$
  - then: $\beta_N > \beta_{N+1} > \cdots > \beta_{2N-2} > \beta_{2N-1} > 1$.

# Tempered Transitions (cont.)

- Starting from the original candidate $m$, at each step we cycle through the $m_i$ as follows:
  - Let $\text{MC}(\mathbf{c}, m)$ denote an MCMC kernel for $\mathbf{c}$ with stationary distribution $m$ where $\tau_1$ is the target density,
  - then we use the following transitions starting at point $\mathbf{c}'_0 = \mathbf{c}^{(t)}$ and iteration $t$ ($\mathbf{N}$ odd):

$$\begin{array}{lr}
\textbf{step 1:} & \mathbf{c}'_1 \sim \text{MC}(\mathbf{c}'_0, m_1) \\
\quad\vdots & \\
\textbf{step N:} & \mathbf{c}'_N \sim \text{MC}(\mathbf{c}'_{N-1}, m_N) \\
\textbf{step N+1:} & \mathbf{c}'_{N+1} \sim \text{MC}(\mathbf{c}'_{N+1}, m_{N-1}) \\
\quad\vdots & \\
\textbf{step 2N-1:} & \mathbf{c}'_{2N-1} \sim \text{MC}(\mathbf{c}'_{2N-2}, m_1).
\end{array}$$

  - Now use a final Metropolis-Hastings decision, accepting $\mathbf{c}'_{2N-1}$ as $\mathbf{c}^{(t+1)}$ with probability:

$$\min\left\{\frac{m_1(\mathbf{c}'_0)}{\tau_1(\mathbf{c}'_0)}\frac{m_2(\mathbf{c}'_1)}{m_1(\mathbf{c}'_1)}\cdots\frac{m_N(\mathbf{c}'_{N-1})}{m_{N-1}(\mathbf{c}'_{N-1})}\cdots\frac{\tau_1(\mathbf{c}'_{2N-1})}{m_1(\mathbf{c}'_{2N-1})}, 1\right\},$$

which preserves the detailed balance condition.

# Tempered Transitions (cont.)

- Substitute the $\beta_i$ parameterization back in for clarity and accept $\mathbf{c}^{(t)}_{1,2N-1}$ with probability:

$$\min\left\{ \left(\frac{m^{1/\beta_1}(\mathbf{c}^{(t)}_0)}{\tau_1(\mathbf{c}^{(t)}_0)}\right) \left(\frac{m^{1/\beta_2}(\mathbf{c}^{(t)}_1)}{m^{1/\beta_1}(\mathbf{c}^{(t)}_1)}\right) \left(\frac{m^{1/\beta_3}(\mathbf{c}^{(t)}_2)}{m^{1/\beta_2}(\mathbf{c}^{(t)}_2)}\right) \dots \right.$$

$$\dots \left(\frac{m^{1/\beta_{N-1}}(\mathbf{c}^{(t)}_{N-2})}{m^{1/\beta_{N-2}}(\mathbf{c}^{(t)}_{N-2})}\right) \left(\frac{m^{1/\beta_N}(\mathbf{c}^{(t)}_{N-1})}{m^{1/\beta_{N-1}}(\mathbf{c}^{(t)}_{N-1})}\right) \left(\frac{m^{1/\beta_{N+1}}(\mathbf{c}^{(t)}_N)}{m^{1/\beta_N}(\mathbf{c}^{(t)}_N)}\right) \dots$$

$$\left. \dots \left(\frac{m^{1/\beta_2}(\mathbf{c}^{(t)}_{2N-3})}{m^{1/\beta_3}(\mathbf{c}^{(t)}_{2N-3})}\right) \left(\frac{m^{1/\beta_1}(\mathbf{c}^{(t)}_{2N-2})}{m^{1/\beta_2}(\mathbf{c}^{(t)}_{2N-2})}\right) \left(\frac{\tau_1(\mathbf{c}^{(t)}_{2N-1})}{m^{1/\beta_1}(\mathbf{c}^{(t)}_{2N-1})}\right), 1 \right\}.$$

- If we notationally "collapse" the ancillary walk it is clear that the standard M-H decision exists:

$$\min\left\{ \left(\frac{m^*(\mathbf{c}^{(t)}_{\upuparrows})}{\tau_1(\mathbf{c}^{(t)}_0)}\right) \left(\frac{\tau_1(\mathbf{c}^{(t)}_{2N-1})}{m^*(\mathbf{c}^{(t)}_{\downdownarrows})}\right), 1 \right\} = \min\left\{ \frac{q(\mathbf{c}|\mathbf{c}')}{q(\mathbf{c}'|\mathbf{c})} \frac{\pi(\mathbf{c}')}{\pi(\mathbf{c})}, 1 \right\}.$$

# Tempered Transitions (cont.)

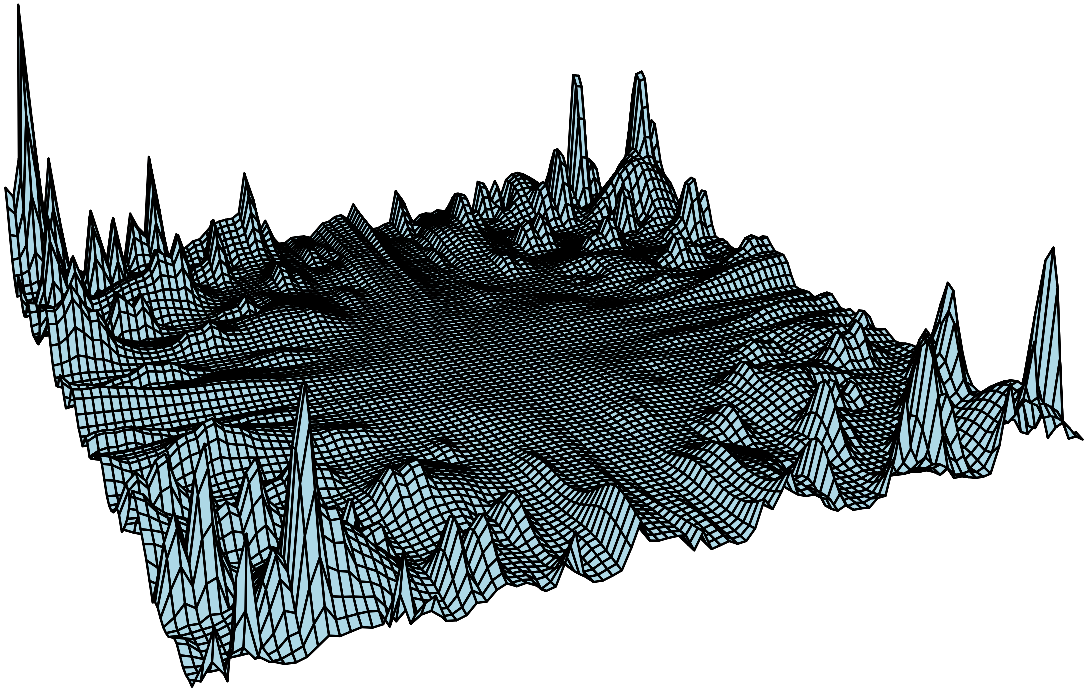
- This sequence of transitions allows excellent exploration of the space, as the density $m_N$ is typically chosen as very "hot", for example, uniform on the entire space.

- Tradeoff: spacing between ladder rungs decision.
  - Setting $\boxed{\beta_i - \beta_{i+1}}$ as small gives higher acceptance rates but poorer mixing.
  - Setting $\boxed{\beta_i - \beta_{i+1}}$ as large is good for mixing around the space but may lead to inordinately high rejection rates.
  - Both criteria can be satisfied with taller ladders (i.e. more steps), for a given difference in $\beta$ values.

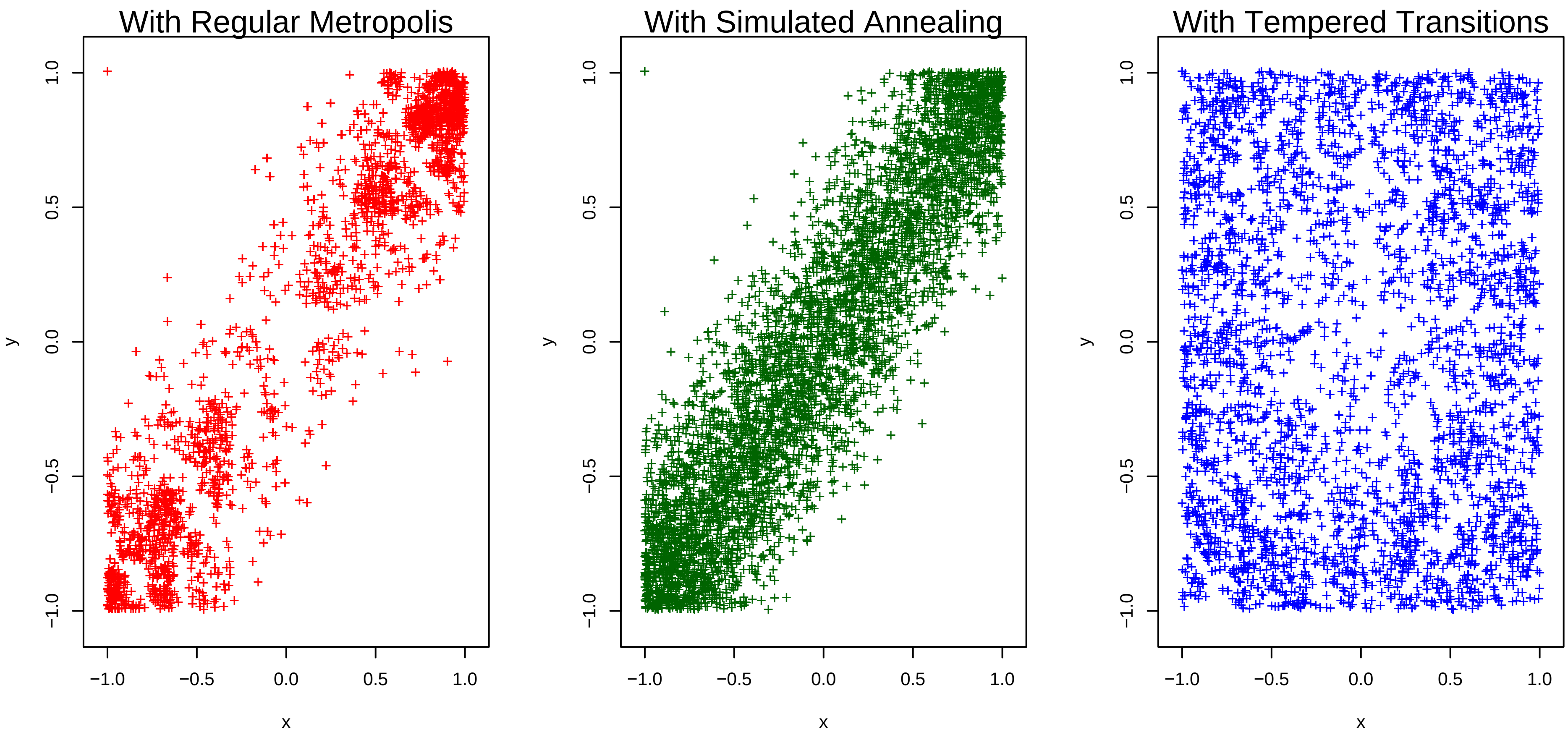
# Comparison of Algorithms

A deliberately "ugly" example objective function on $[-1,1]^2$:

$f(x,y) = \text{abs}((x\sin(20y-90) - y\cos(20x+45))^3 a\cos(\sin(90y+42)x) + (x\cos(10y+10) - y\sin(10x+15))^2 a\cos(\cos(10x+24)y))$.



# Comparison of Algorithms (5,000 iterations only)



## (New) Alternative 4 Dynamic Tempered Transitions

- Ladder height issue:
  - When the area around the chain is highly irregular, it is better to have a lower (cooler maximum temperature) ladder in order to better explore modes.
  - When the area around the chain is smooth, it is better to have a longer (hotter maximum temperature) ladder in order to quickly traverse the low density region.
- Number of ladder rungs issue:
  - Setting the number for a given ladder height as too small can reduce the acceptance rate because high quality candidates may not be offered.
  - Conversely, settings the number as too large can also reduce the acceptance rate because of the product in the Metropolis-Hastings decision step.

# Dynamic Tempered Transitions (cont.)

- Our strategy:
  - Specify a *distribution* of ladders all having the same number of rungs, but differing heights (differing maximum temperatures).
  - Observe the multidimensional curvature at the current Markov chain location and specify a greater probability of selecting a cooler ladder when this curvature is high, and a greater probability of selecting a hotter ladder when the curvature is low.
  - We treat the number of rungs as a nuisance parameter which can be fixed at the beginning of the chain or tuned during the early runs by comparing acceptance probabilities.

# Dynamic Tempered Transitions Details

- Define $\lambda$ to be a discrete parameter that indexes ladders.

- For each $\mathbf{c}$, let $\rho(\lambda|\mathbf{c})$ be a proper conditional probability distribution, that is, $\rho(\lambda|\mathbf{c}) \geq 0$ and $\int \rho(\lambda|\mathbf{c})d\lambda = 1$.

- Next, let $m_\lambda(\mathbf{c}, \mathbf{c}')$ be a temperature indexed distribution similar to the described tempered transitions ladder according to Neal (1996).

- We want to take a mixture of these ladders now. Therefore our candidate generating distribution becomes:

  $$g_\lambda(\mathbf{c}'|\mathbf{c}) = m_\lambda(\mathbf{c}, \mathbf{c}')\rho(\lambda|\mathbf{c}),$$

  and form a Metropolis kernel based on $g_\lambda(\mathbf{c}'|\mathbf{c})$ and $f(\mathbf{c})$.

# Dynamic Tempered Transitions Details (cont.)

- As an example, suppose that there are $i = 1, \ldots, k$ ladders, and as $i$ increases the ladders get hotter.

- If $|f''(\mathbf{c})|$ is big (so we are near a mode) we might want to favor the cooler ladders.

- To do this we can take $\rho(\lambda|\mathbf{c})$ to be a binomial mass function with $k$ trials and success probability $p(\mathbf{c})$, where

$$\text{logit}[p(\mathbf{c})] = a - b|f''(\mathbf{c})|, \quad b > 0.$$

- So big values of $|f''(\mathbf{c})|$ would result in small $p(\mathbf{c})$, which would favor the smaller values of $i$ and the cooler ladders.

# Dynamic Tempered Transitions Details (cont.)

- The key challenge is that by making the behavior the Markov chain adjust to its surroundings (i.e. conditional on $\mathbf{c}$), we run the risk of losing the detailed balance equation.

- Let $f(\mathbf{c})$ be the stationary distribution (objective function), let $g_\lambda(\mathbf{c}'|\mathbf{c})$ be a candidate distribution, and let $MC_\lambda(\mathbf{c},\mathbf{c}')$ be the associated transition kernel.

- By the construction of the Metropolis algorithm, $MC_\lambda(\mathbf{c}',\mathbf{c})$ is now given by:

$$MC_\lambda(\mathbf{c}',\mathbf{c}) = \min\left\{\frac{f(\mathbf{c}')g_\lambda(\mathbf{c}|\mathbf{c}')}{f(\mathbf{c})g_\lambda(\mathbf{c}'|\mathbf{c})}, 1\right\} g_\lambda(\mathbf{c}'|\mathbf{c}) + (1-r(\mathbf{c}))\delta_{\mathbf{c}}(\mathbf{c}'),$$

  where:

$$r(\mathbf{c}) = \int \min\left\{\frac{f(\mathbf{c}')g_\lambda(\mathbf{c}|\mathbf{c}')}{f(\mathbf{c})g_\lambda(\mathbf{c}'|\mathbf{c})}, 1\right\} g_\lambda(\mathbf{c}'|\mathbf{c})d\mathbf{c}'$$

  and $\delta_{\mathbf{c}}(\mathbf{c}') = 1$ if $\mathbf{c} = \mathbf{c}'$ and zero otherwise.

# Dynamic Tempered Transitions Details (cont.)

- The kernel $MC_\lambda(\mathbf{c},\mathbf{c}')$ now satisfies detailed balance with $f(\mathbf{c})$ as the stationary distribution *at each individual step* (exactly from Robert and Casella 1999, Theorem 6.2.3, with proof), so the full Markov chain kernel is interpreted as a (very high dimension) mixture kernel.

- Because:

$$\min\left\{\frac{f(\mathbf{c}')g_\lambda(\mathbf{c}|\mathbf{c}')}{f(\mathbf{c})g_\lambda(\mathbf{c}'|\mathbf{c})},1\right\}g_\lambda(\mathbf{c}'|\mathbf{c})f(\mathbf{c}) = \min\left\{\frac{f(\mathbf{c})g_\lambda(\mathbf{c}'|\mathbf{c})}{f(\mathbf{c}')g_\lambda(\mathbf{c}|\mathbf{c}')},1\right\}g_\lambda(\mathbf{c}|\mathbf{c}')f(\mathbf{c}')$$

  and:

$$(1-r(\mathbf{c}))\delta_{\mathbf{c}}(\mathbf{c}')f(\mathbf{c}) = (1-r(\mathbf{c}'))\delta'_{\mathbf{c}}(\mathbf{c})f(\mathbf{c}')$$

  for each selected $\lambda$.

# Example: A Probabilistic Spatial Model of Voting Under Uncertainty

- $J$ candidates, $N$ voters, and a compact, convex $K$-dimensional Euclidean issue-space $\mathbb{S}^K$.

- The voter is assumed to vote for the candidate who has a $K$-dimensional position the closest to this voter's ideal point: *sincere proximity voting.*

- Our hypothetical Candidate picks a point in $\mathbb{S}^K$ designed to maximize her expected votes, given the other candidates' position.

- The position of candidate $j$ $(j = 1, \ldots, J)$ is the $K$-length vector:

$$\mathbf{c}_j = [C_{j1}, C_{j2}, \ldots, C_{jK}].$$

- Voter $i$'s *ideal point* in $\mathbb{S}^K$ is the $K$-length vector:

$$\mathbf{v}_i = [V_{i1}, V_{i2}, \ldots, V_{iK}].$$

# Example: A Probabilistic Spatial Model of Voting Under Uncertainty (cont.)

- The utility of *candidate* $j$ to voter $i$ is the negative (vector) distance between $\mathbf{c}_j$ and $\mathbf{v}_i$, plus a zero-mean uncertainty term independent across candidates($\mathbf{E}_j$):

$$\mathbf{U}_{ij} = \mathbf{E}_j - \mathbf{D}_{ij} = \mathbf{E}_j - ||\mathbf{v}_i - \mathbf{c}_j||$$

  where $||.||$ denotes the vector $K$-norm so $\mathbf{D}_{ij}$ is squared Euclidean distance.

- Voter $i$ prefers *candidate* $j$ over *candidate* $\ell$ if her utility for $j$ exceeds her utility for $\ell$:

$$\mathbf{U}_{ij} - \mathbf{U}_{i\ell} = (\mathbf{E}_j - \mathbf{D}_{ij}) - (\mathbf{E}_\ell - \mathbf{D}_{i\ell}) > 0,$$

  defining

$$\boxed{\mathbf{E}_{j\ell} = \mathbf{E}_\ell - \mathbf{E}_j} \quad \text{and} \quad \boxed{\mathbf{D}_{ij,\ell} = \mathbf{D}_{i\ell} - \mathbf{D}_{ij}}$$

  so

$$\mathbf{U}_{ij} - \mathbf{U}_{i\ell} = \mathbf{D}_{ij,\ell} - \mathbf{E}_{j\ell}$$

# Example: A Probabilistic Spatial Model of Voting Under Uncertainty (cont.)

- Voter $i$ votes for candidate $j$ over *all* others if:

$$\mathbf{E}_{j\ell} < \mathbf{D}_{ij,\ell}, \ \ell = 1, 2, \ldots, j-1, j+1, \ldots, J.$$

- Of course voters are assumed to be comparing *all* candidates, so for $j$ as the baseline comparison candidate, the $(J-1)$-length uncertainty vector can be treated as multivariate normal:

$$\mathbf{e}_j = \left[\mathbf{E}_{j1}, \ldots, \mathbf{E}_{j(j-1)}, \mathbf{E}_{j,(j+1)}, \ldots, \mathbf{E}_{jJ}\right] \sim \phi(\mathbf{0}, \boldsymbol{\Delta}_j).$$

- Collect the $K$-dimensional $J-1$ distance cross-candidate comparisons to *candidate* $j$ into a single vector:

$$\mathbf{d}_j = \left[\mathbf{D}_{ij,1}, \ldots, \mathbf{D}_{ij,(j-1)}, \mathbf{D}_{ij,(j+1)}, \ldots, \mathbf{D}_{ij,J}\right].$$

# Example: A Probabilistic Spatial Model of Voting Under Uncertainty (cont.)

- So the probability that voter $i$ votes for *candidate* $j$ is the CDF of the multivariate normal at the (vector-valued) point $\mathbf{d}_j$:

$$\begin{aligned} P(i,j) &= P(\mathbf{E}_{j\ell} < \mathbf{D}_{ij,\ell}, \ell = 1,\ldots,j-1,j+1,\ldots,J) \\ &= \int_{-\infty}^{\mathbf{D}_{ij,1}} \cdots \int_{-\infty}^{\mathbf{D}_{ij,j-1}} \int_{-\infty}^{\mathbf{D}_{ij,j+1}} \cdots \int_{-\infty}^{\mathbf{D}_{ij,J}} \phi(\mathbf{0}, \boldsymbol{\Delta}_j) d\mathbf{E}_{j,1} \cdots d\mathbf{E}_{j,j-1} d\mathbf{E}_{j,j+1} \cdots d\mathbf{E}_{j,J}. \end{aligned}$$

- This setup makes it possible to calculate the expected vote totals for the $j$th candidate:

$$\tau_j(\mathbf{c}|\mathbf{v}) = EV_j(\mathbf{c}) = \sum_{i=1}^{N} P(i,j)$$

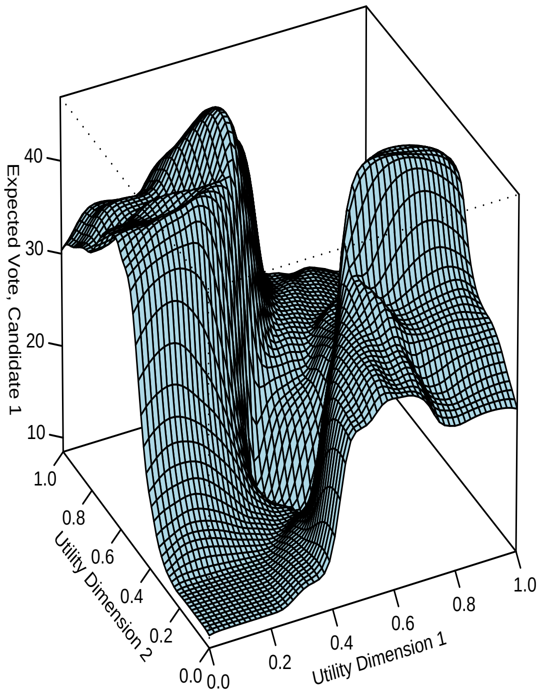
  and therefore every candidate.

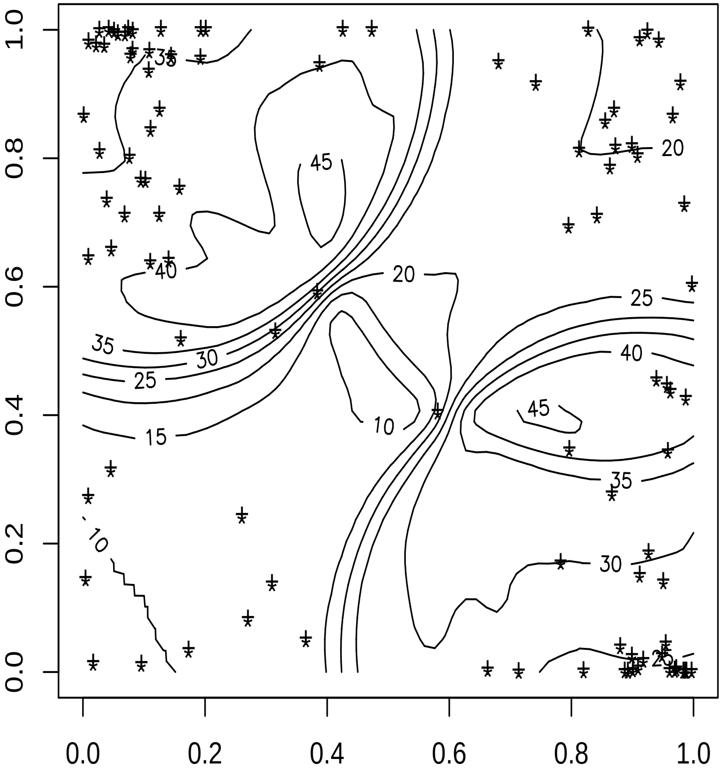
- We will restrict ourselves to the case of *fixed* competing candidates and the determination of the *best reply surface* for a hypothetical candidate of interest.

# Small Example

- 3 candidates, 100 voters, over a standardized 2-dimensional issue-metric, where we evaluate the expected number of votes for *candidate 1* taking all possible issue positions.

- Voter ideal points are drawn from a mixture of beta densities to reflect some division of preferences across two roughly defined groups, and $\mathbf{e}_j \sim \mathcal{N}(0, 0.02)\ \forall j$.





# Application to Some Real Data

- American National Election Study (ANES) of the 2000 Presidential election.
- 1462 potential voters surveyed prior to the election.
- We analyze 10 policy dimensions where respondents place themselves, to produce: $\mathbf{v}_i = [V_{i1}, V_{i2}, \ldots, V_{iK}]$.
- Policy issues given by nominal scales:
  - political ideology
  - preference on government spending
  - preference on defense spending
  - government should help generate jobs
  - government should help African Americans economically
  - abortion
  - environment vs. industrial development
  - guns
  - role of women in society
  - increased/decreased regulation

## Application to Some Real Data

- The positions for Gore and Bush are modal evaluations from respondents across ten dimensions:

$$\mathbf{c}_{Gore} = [C_{Gore,1}, C_{Gore,2}, \ldots, C_{Gore,10}]$$
$$\mathbf{c}_{Bush} = [C_{Bush,1}, C_{Bush,2}, \ldots, C_{Bush,10}].$$

- Our question: can we find a third candidate position that beats both on these policy issues?

$$\tau(\mathbf{c}_{new}|\mathbf{v}) = \sum_{i=1}^{2} P(i, new) > \tau(\mathbf{c}_{Gore}|\mathbf{v}) \text{ and } \tau(\mathbf{c}_{Bush}|\mathbf{v})$$

# Application to Some Real Data

- What does this mean electorally?

- We compare $\mathbf{c}_{Gore}$ and $\mathbf{c}_{Bush}$ to the greatest posterior mode:

| ideology | spending | defense | jobs | blacks | abortion | environment | guns | gender | regulation |
|---|---|---|---|---|---|---|---|---|---|
| 3.2423 | 3.2478 | 3.2384 | 3.2195 | 3.2230 | 3.2494 | 3.2388 | 3.2412 | 2.6322 | 3.0006 |

using

$$\tau_j(\mathbf{c}|\mathbf{v}) = EV_j(\mathbf{c}) = \sum_{i=1}^{N} P(i,j)$$

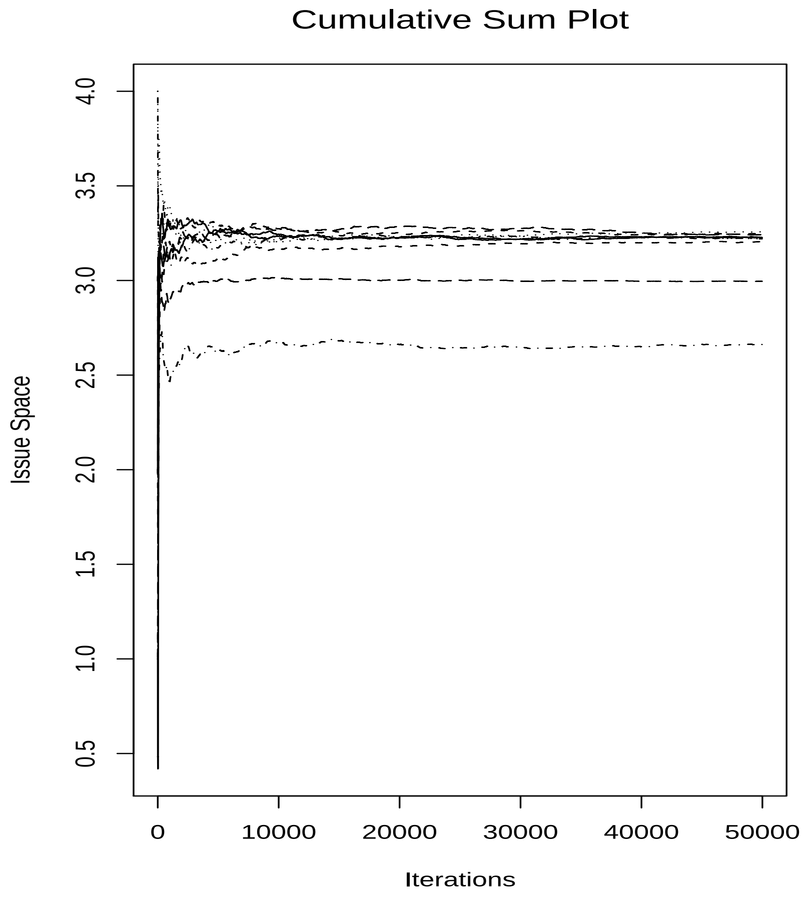
- Results produced using SIR to get tail values from: $\sum_{i=1}^{N} \int\limits_{-\infty}^{\mathbf{D}_{ij,1}} \cdots \int\limits_{-\infty}^{\mathbf{D}_{ij,j-1}} \int\limits_{-\infty}^{\mathbf{D}_{ij,j+1}} \cdots \int\limits_{-\infty}^{\mathbf{D}_{ij,J}} \phi(\mathbf{0}, \boldsymbol{\Delta}_j) d\mathbf{E}_{j,1} \cdots d\mathbf{E}_{j,j-1} d\mathbf{E}_{j,j+1} \cdots d\mathbf{E}_{j,J}$

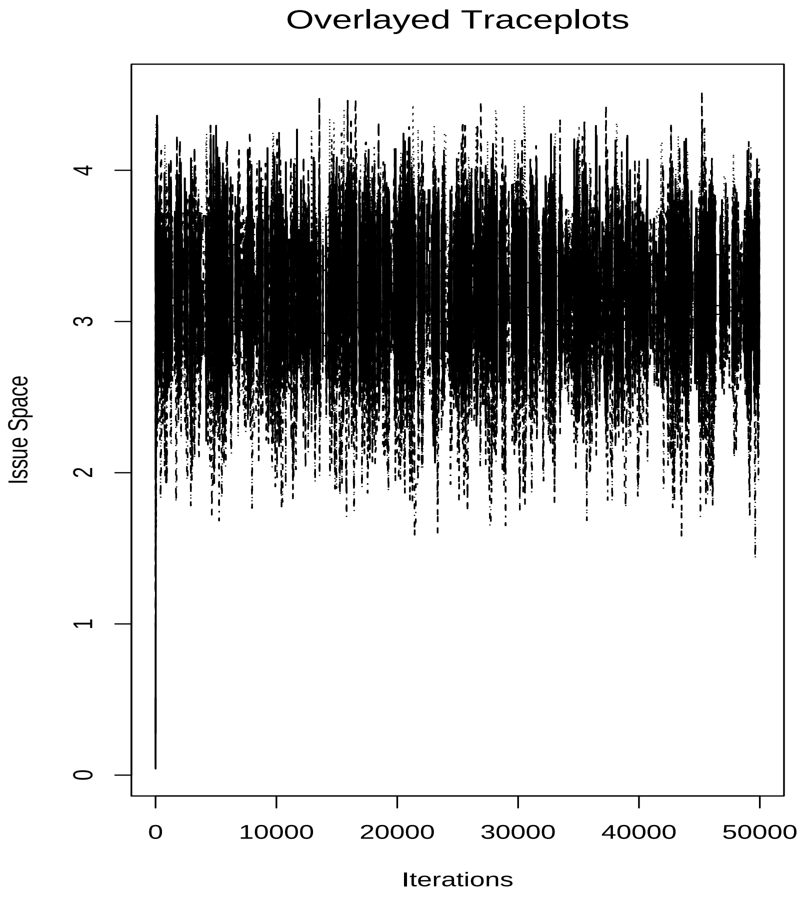
$$\tau_j(Gore|\mathbf{v}) = 0.34758$$

$$\tau_j(Bush|\mathbf{v}) = 0.17477$$

$$\tau_j(Candidate|\mathbf{v}) = 0.47100$$

# Application to Some Real Data





# Random (anonymous) Testimonials

- "…the various version of choice based sampling maximum likelihood all suffer a severe computational drawback. Each requires numerous evaluations of integrals of the form $\int_Z P(i,z,\theta)p(z)dz$."

- "…solving $M$ different $k$-dimensional first-order conditions is a maximization problem of great complexity which leads to them use a computerized search to locate locally optimal candidate locations in a 2-dimensional 6-candidate election."

- "We managed to optimize the criterion with a simplex-type algorithm, but the convergence process took a very long time. The estimated model was, however, relatively simple, and it is therefore doubtful whether this type of numerical algorithm is well suited for more complicated econometric models (such as models with explanatory variables)."