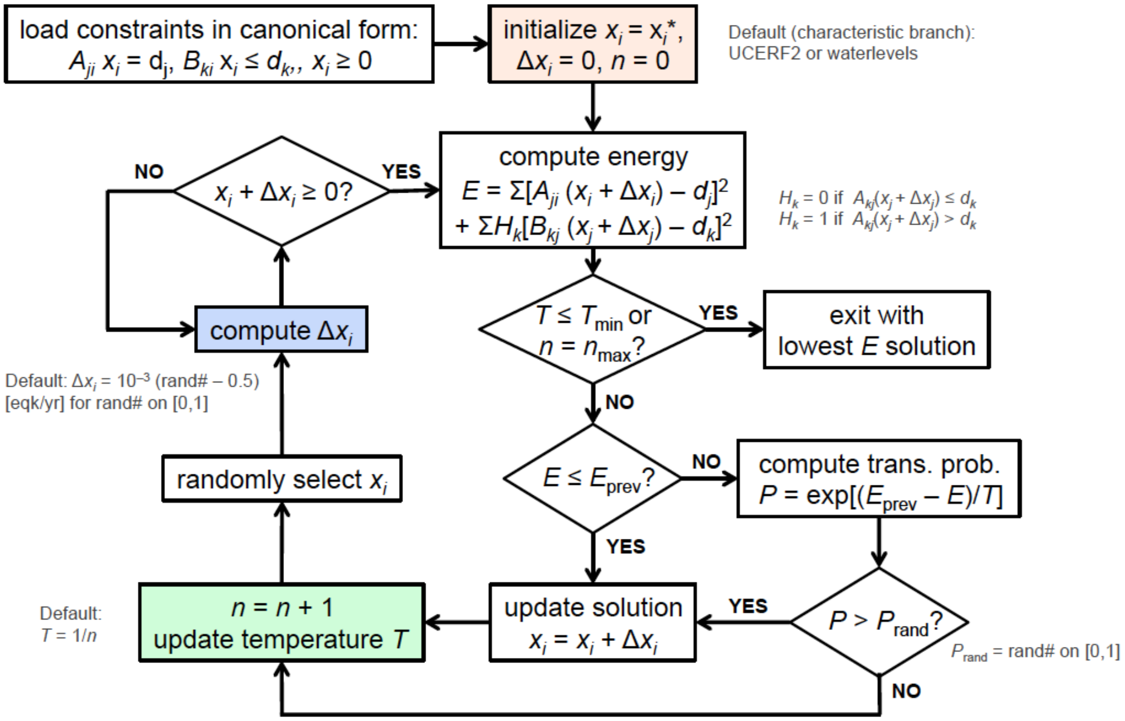
Review of the Grand Inversion
by T. H. Jordan (12/18/12)

Application of the Gran Inversion (GI) to the derivation of UCERF3.1 is described in the draft report, "Proposed Time-Independent Uniform California Earthquake Rupture Forecast, Version 3.1 (UCERF3.1)" by Field et al. (Nov, 2012), and in Appendix N, "Grand Inversion Implementation and Exploration of Logic-Tree Branches" by Page et al. (Nov 2012). On Dec 14, Morgan Page and I met to review the Grand Inversion (GI) codes, computational procedures, and UCERF3.1 results. This written review summarizes the findings and recommendations that came out of our discussions.

1. Algorithm and Code Types. There are two types of code used in the Grand Inversion, a single-processor (serial) code and a multiple-processor (parallel) code. The serial code uses a fairly standard implementation of simulated annealing (SA), as described in the following flowchart:

Grand Inversion simulated annealing algorithm for UCERF3.1
(Page, Field & Milner, 2012)



load constraints in canonical form: $A_{ji}\, x_i = d_j$, $B_{ki}\, x_i \le d_k$, $x_i \ge 0$
initialize $x_i = x_i^*$, $\Delta x_i = 0$, $n = 0$
Default (characteristic branch): UCERF2 or waterlevels
compute energy $E = \Sigma[A_{ji}\,(x_i + \Delta x_i) - d_j]^2 + \Sigma H_k[B_{kj}\,(x_j + \Delta x_j) - d_k]^2$
$H_k = 0$ if $A_{kj}(x_j + \Delta x_j) \le d_k$
$H_k = 1$ if $A_{kj}(x_j + \Delta x_j) > d_k$
NO
$x_i + \Delta x_i \ge 0$?
YES
compute Δx_i
Default: $\Delta x_i = 10^{-3}$ (rand# – 0.5) [eqk/yr] for rand# on [0,1]
$T \le T_{min}$ or $n = n_{max}$?
YES
exit with lowest E solution
NO
$E \le E_{prev}$?
NO
compute trans. prob. $P = \exp[(E_{prev} - E)/T]$
YES
randomly select x_i
update solution $x_i = x_i + \Delta x_i$
YES
$P > P_{rand}$?
P_{rand} = rand# on [0,1]
NO
$n = n + 1$ update temperature T
Default: $T = 1/n$

The data constraints are weighted sets of linear equations and inequalities. The components of the solution vector—the rupture rates [yr^{-1}]—are also constrained to have minimum non-negative values ("waterlevels"). Prior to inversion, the constraint sets are reduced to a canonical form: $\{A_{ji}\, x_i = d_i : j = 1, \ldots, J\}$, $\{B_{ki}\, x_i \leq d_k : k = 1, \ldots, K\}$, $\{x_i \geq 0 : i = 1, \ldots, M\}$.

The solution vector is initialized and then randomly perturbed one component at a time. If the perturbed component is positive, the energy E (squared misfit) is calculated. If the transition probability $P = \min\{1, \exp[(E_{\text{prev}} - E)/T]\}$ equals or exceeds a random value on [0,1], the solution is updated. According to this criterion, a perturbation that decreases the solution energy is always kept, whereas one that increases the energy is kept with a frequency that decreases with the temperature T. In either case, the iteration number n is incremented, and T is decreased.

Once the constraints are fixed, the main controls on the inversion pathway are the solution-vector initialization (pink box in flowchart), the amplitude of the model perturbation Δx_i (blue box), and the decrease in temperature T with iteration number n (green box). The default values are annotated on the flowchart.

In the parallel version, the SA process is initiated as 4 threads on each of up to 50 nodes; after some run time (default: $n = 200$), the results are compared, and all threads are restarted from the solution with the lowest energy. Comparing solutions and selecting the best-fit individual solution for reproduction deviates from a standard SA algorithm by augmenting it with a step used in some types of genetic algorithms. This augmented procedure accelerates the annealing schedule relative to the serial algorithm, perhaps reducing access to parts of the solution manifold. Kevin Milner has optimized the parallel version to run on USC and TACC supercomputers. The results were presented as a poster at the 2011 SCEC Annual Meeting and are briefly discussed in Appendix N.

- *Documentation*: The documentation of the serial and parallel SA algorithms in Appendix N should include the level of detail in the flowchart. Any differences with standard SA practices should be fully described, and their possible effects on the solution sets should be assessed.

2. Solution Comparison and Reproducibility. Owing to the stochastic features of the SA algorithm, the annealing pathway is not a deterministic function of the input data; any solution vector **x** obtained after a finite number of iterations is not unique, and it cannot be exactly reproduced by re-running the code. In the Grand Inversion for UCERF3.1, there are $M = 234{,}188$ unknowns in the solution vector, and $N = J + K = 34{,}466$ equations and inequalities in the design matrix. Because $M \gg N$, the manifold of acceptable solutions is expected to be quite large. The nature of this nonuniqueness cannot be evaluated using standard linear resolving power tests owing to the (nonlinear) inequality constraints.

In the current SA procedure, solution vectors are generated from 100 runs with the same starting model and annealing schedule and averaged to get a mean vector, which is also a solution. The ranges of individual vector components in this solution set are valuable measures of nonuniqueness, even though they are statistically incomplete. (Sampling the full covariation among all model parameters would presumably require hundreds of thousands of runs.)

These considerations suggest testing the reproducibility of the solution set by inverting synthetic data.

- *Synthetic data tests*: Invert synthetic data calculated from the mean vector of the solution set derived for the UCERF3.1 Zeng characteristic reference branch. The synthetic data should include the mean-solution MFDs as targets (rather than UCERF2 MFDs).

This experiment will evaluate how well a mutually consistent set of constraints can be satisfied by the SA algorithms and how well a data-generating model can be reproduced. Example questions to be answered should include: Does the data-generating model have components that fall frequently enough within a specified quantile range of the 100-run synthetic data inversion (e.g., 90% within the 5%-95% range)? If not, how big does the solution set have to be to contain the data-generating model within a particular quantile range? Does this confidence-range criterion provide an adequate definition of "statistical reproducibility"?

3. Code Verification. Both the serial and parallel SA algorithms are very simple and easily implemented in Java, which makes line-by-line code verification easier, though by no means error-proof. The serial version has been tested for convergence using simple models and data sets. The parallel version has been cross-verified with the serial version, though these tests have not been documented. In particular, it is unclear how well solution vectors from the selection-based parallel code agree with those from the SA-only serial code.

- *Documentation*: Verification and reproducibility exercises should include quantitative comparisons of the solution sets from the serial and parallel codes, and the results should be documented in Appendix N.

4. Evaluating Nonuniqueness. A major issue is whether the nonuniqueness is "local"; i.e., whether the run-to-run variations in the rupture rates are nearly the same in terms of hazard measures that integrate over the rupture set. During the review, we examined the rupture rates on different fault sections, including both isolated faults (e.g., Battle Creek) and well-connected faults (e.g., S. Mojave section of SAF). As required by the averaging over the solution set, individual solutions are "less smooth and more compact" than the mean solution; i.e., higher rates are assigned to fewer ruptures. The solution-to-solution variations in the participation MFDs were substantial in some cases, but the data fits appeared to be comparable. These observations should be quantified.

- *Energy convergence tests*: *E* vs. *n* plots should be monitored for the 100-run reference-branch solution sets to see how well solutions from the serial and parallel codes converge to similar data fits using the preferred annealing schedules. The mean and variance in the misfit (or the more robust median and median-absolute-deviation) should be computed from the solution set for all data points, and the hazard implications of data misfit outliers should be evaluated.

For the Zeng characteristic reference branch, the slip-rate variations that we examined appeared to be small, even at the fault-element scale. These observations, together with the very small solution-set variations in the hazard curves computed by Peter Powers, suggest that the primary nonuniqueness is local. This hypothesis should be tested by more systematic analyses of slip-rate variations and related hazard variations.

- *Slip-rate convergence tests*: The solution-set range in slip rates for all fault elements should be analyzed, and the hazard implications of elements with high variance should be evaluated.

As a further test of localization, non-local tradeoffs in the solution set should be explored by looking for large cross-correlations among rupture rates on different fault segments.

A second issue is the dependence of the solution-set range on the starting model. For the characteristic branches, the current starting model uses UCERF2 rates (plus waterlevels where needed). Early testing by Page involved different starting models, including a near-zero (waterlevel) model, but these tests should be repeated for the UCERF3.1 configuration and expanded to starting models with randomized components.

- *Starting-model convergence tests*: Reference-branch solution sets should be computed for various types of starting models, including UCERF2, near-zero, and randomized starting models, and the solution-set ranges should be examined for hazard implications.

The third issue is the dependence of the solution set on the annealing schedule, as controlled in the green and blue boxes of the flowchart. In the default algorithm, the temperature *T* decreases inversely with the iteration number, which is a fast annealing schedule, and the model perturbation is independent of temperature and the model value. As noted in Appendix N, SA convergence to a global minimum energy is guaranteed only for logarithmically slow annealing schedules, and standard SA practice scales the model perturbation with *T*.

- *Annealing-schedule convergence tests*: Reference-branch solution sets should be computed for slower annealing schedules, and differences in the solution-set ranges from the fast-annealed reference branch should be examined for hazard implications. They should also be computed using different perturbation schemes, including ones in which the perturbation amplitude decreases with *T*.

5. Other Issues. According to the expert opinion of some reviewers, the UCERF3.1 reference branch solutions do not provide an adequate fit to the data. One issue is

whether these inadequacies can be attributed to deficiencies in the inversion algorithm, or whether they signal inconsistencies in the competing data sets. We examined one case—the S. Mojave section of the San Andreas fault—where both the paleoseismic event rates and UCERF2 MFD are under-predicted by the Zeng reference branch. A series of tests suggested that the inconsistency is data-related; e.g., the fit to the paleoseismic rates trades off strongly with fit to the regional MFD for Southern California. The inconsistency could result from over-estimation of the off-fault seismicity (which is taken off the top before inversion), improper weighting of the relative slip rates on low-rate faults, or slip rates on the SAF biased too low by the geodetic data. Further testing should investigate these possibilities.

6. Conclusions.

a. The GI seems to be doing what it's supposed to do, although further verification steps are recommended. In particular, the dependence of the solutions on starting model and annealing schedule need to be investigated for the full UCERF3.1 problem configuration.

b. The SA algorithm used in the parallel code differs from that used in the serial code. Comparison tests are recommended to make sure the effects of the algorithmic differences on the solutions are understood.

c. For UCERF3.1, the number of constraints is substantially less than the number of model components, so the solution is nonunique. The manifold of acceptable solutions can be partially explored through large suites of inversions.

d. Averaging over large sets of solutions provides a smoother, less-compact solution vector, suitable for use as a reference solution. The variation of rupture rates and integral measures such as slip rates from large solution sets should be used to quantify the nonuniqueness and assess its hazard implications.

e. The nonuniqueness of UCERF3.1 solutions appears to be primarily local; e.g., integration over the rupture rates for different solutions yields similar slip rates, even at the level of individual fault elements. To test this hypothesis, non-local tradeoffs in the solution set should be explored by looking for large cross-correlations among rupture rates on different fault segments.

f. Owing to the stochastic nature of the SA algorithm, GI solutions are not deterministically reproducible. Given the importance of reproducibility in hazard calculations, it may be useful to define "statistical reproducibility" using location and dispersion measures from large solution sets. The validity of this concept should be investigated using synthetic-data and real-data tests.

g. Solutions for the UCERF3.1 Zeng characteristic reference branch show problems in fits to data subsets, but these misfits appear to result from inconsistencies in the data, rather than problems in the data inversion.