BFGS vs L-BFGS-B Diffusion Failure Study

Automated Benchmark Report

February 11, 2026

What We Ran

Goal

This study isolates diffusion-coefficient inference and compares `BFGS` against `L-BFGS-B`, with focus on forward-solve failures.

Setup

Problem: `infer_D` only. Methods: `BFGS`, `L-BFGS-B`. Noise levels: $\sigma \in \{0, 0.005, 0.02\}$. Seeds used for $\sigma = 0$: [20270400]; seeds used for $\sigma > 0$: [20270401, 20270402, 20270403, 20270404, 20270405, 20270406, 20270407, 20270408, 20270409, 20270410]. Total runs: 168; failed runs: 8.

How to read the plots

Each title and colorbar states whether higher or lower values are preferred.

Method Comparisons

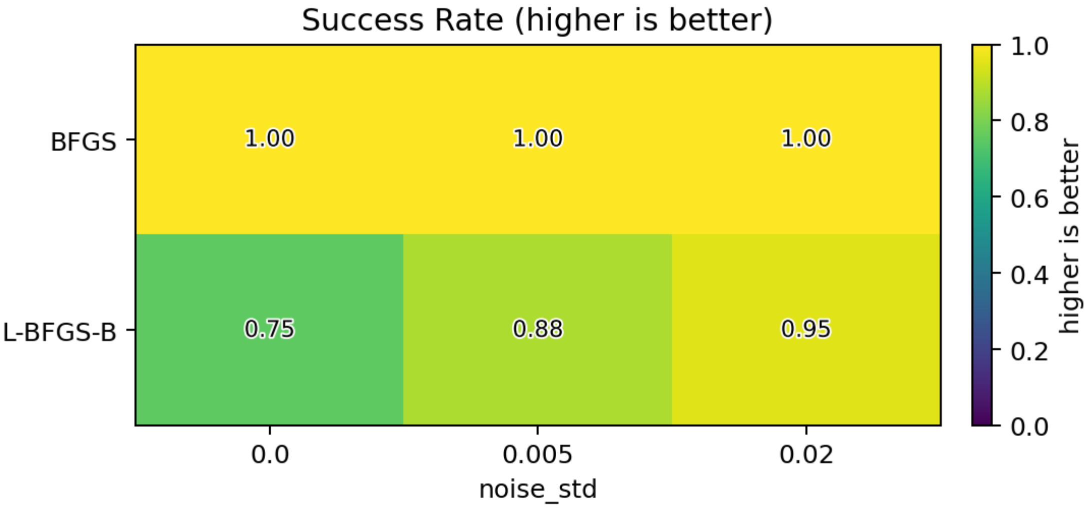

Figure 1: Success rate by method/noise (higher is better).

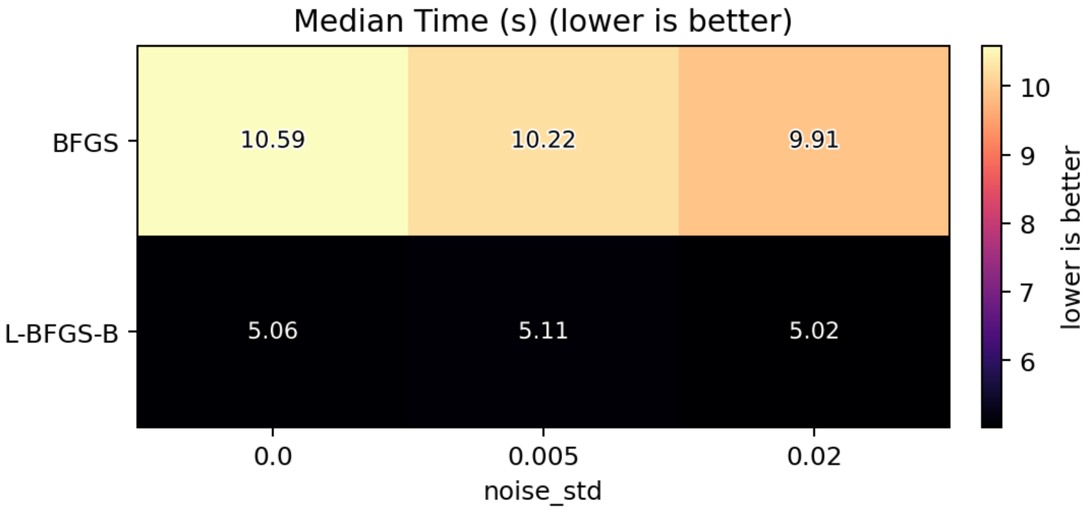

Figure 2: Median wall-clock time (s) for successful runs (lower is better).

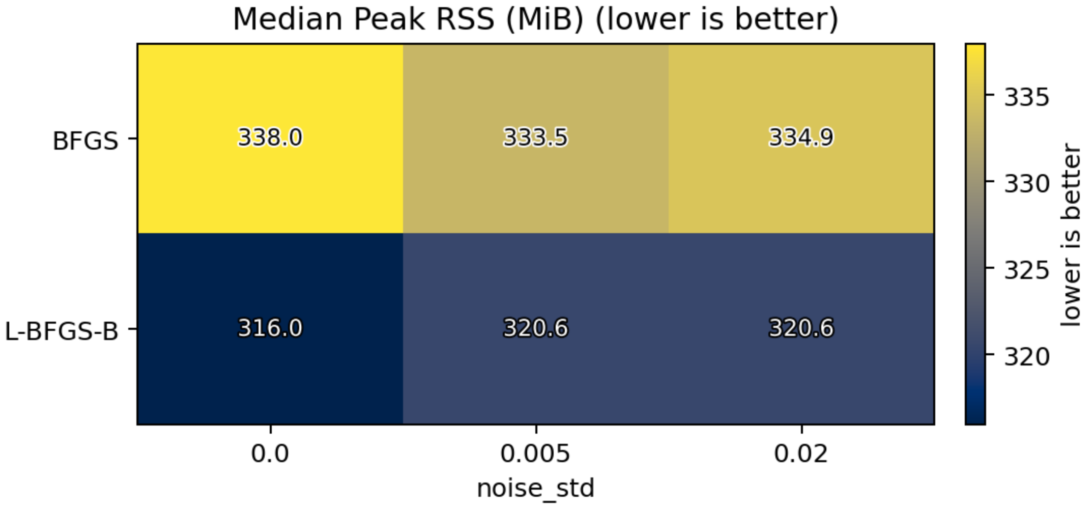

Figure 3: Median peak RSS memory (MiB) for successful runs (lower is better).

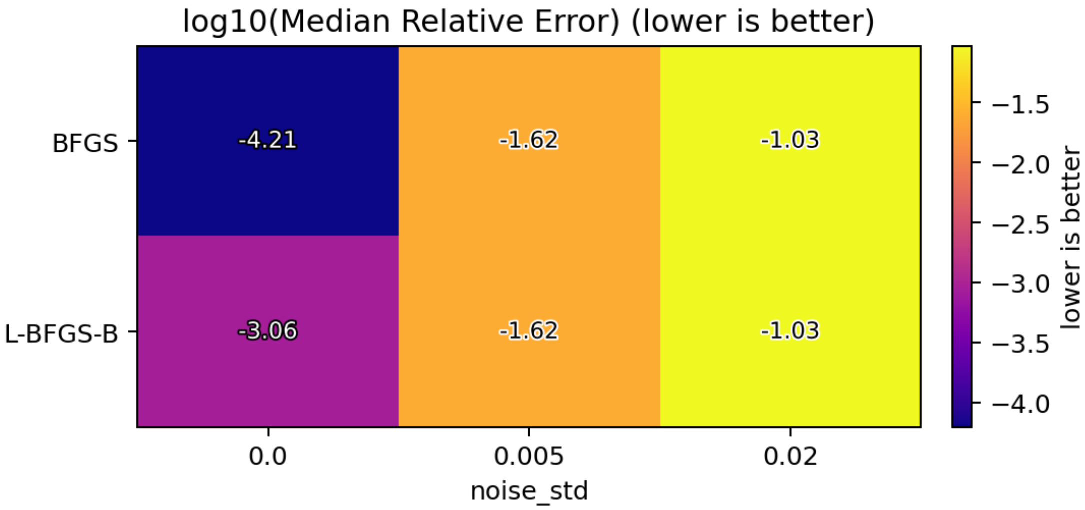

Figure 4: $\log_{10}$ median relative error by method/noise (lower is better).

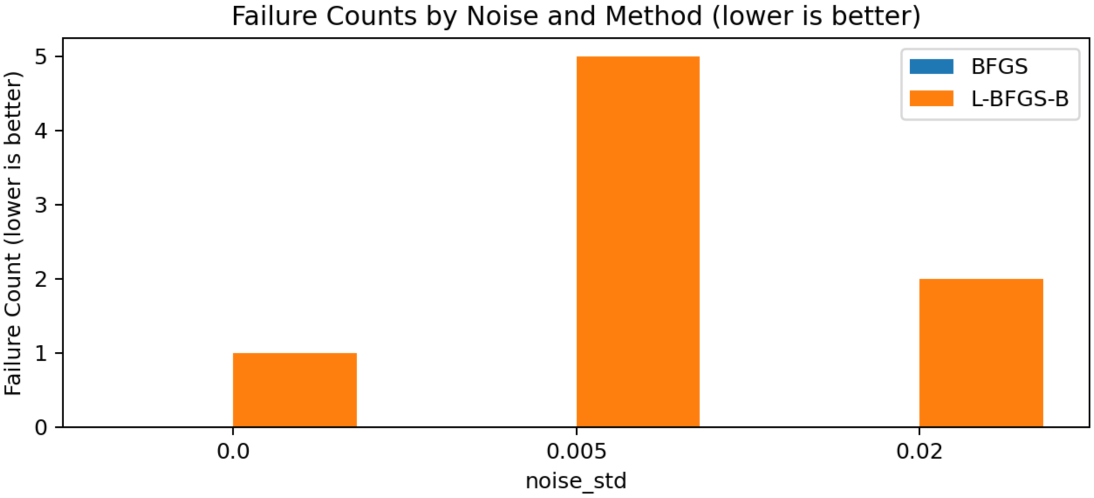

Figure 5: Failure counts by noise and method (lower is better).

Aggregated Results Across Noise Levels

Method	Success (%)	Median Time (s)	Median RSS (MiB)	Median RelErr
BFGS (σ=0)	100	10.6	338	6.21e-05
BFGS (σ=0.005)	100	10.2	334	0.0242
BFGS (σ=0.02)	100	9.91	335	0.0925
L-BFGS-B (σ=0)	75	5.06	316	0.000875
L-BFGS-B (σ=0.005)	87.5	5.11	321	0.0241
L-BFGS-B (σ=0.02)	95	5.02	321	0.094

Crash Analysis

What failed?

All 8 failures were in `L-BFGS-B` runs; `BFGS` had 0 failures. Dominant reason: `DIVERGED_LINE_SEARCH` (8 of 8 failures).

Across-Seed Stability for Noisy Cases ($\sigma > 0$)

Method	Failures / Runs	Failure Rate (%)	Seeds with failures
BFGS	0/80	0	none
L-BFGS-B	7/80	8.75	[20270401, 20270402, 20270404, 20270406, 20270407, 20270409, 20270410]

Failure-Causing D Patterns

Failure initial guesses: {'[10.0, 10.0]': 8}; failure true-D cases: {'[1.0, 3.0]': 8}. Logged min-D component in failed runs: min=0.0323, median=0.23, max=0.239. Logged min-D component in successful runs: min=1.11e-19, median=0.3, max=0.394.

σ	Seed	D_{true}	D_0	Min logged trial D	Last logged D
0	20270400	[1.0, 3.0]	[10.0, 10.0]	[0.23360679, 0.36898834]	[1.89295046, 2.31792262]
0.005	20270401	[1.0, 3.0]	[10.0, 10.0]	[0.2348071, 0.36684491]	[1.90479801, 2.32004539]
0.005	20270402	[1.0, 3.0]	[10.0, 10.0]	[0.03229418, 1.01357971]	[1.42543256, 2.19932283]
0.005	20270404	[1.0, 3.0]	[10.0, 10.0]	[0.23570039, 0.36526913]	[1.91615451, 2.32437545]
0.005	20270406	[1.0, 3.0]	[10.0, 10.0]	[0.04335572, 1.14466137]	[1.41816252, 2.20680239]
0.005	20270409	[1.0, 3.0]	[10.0, 10.0]	[0.04891558, 1.19075142]	[1.42703783, 2.2072424]
0.02	20270407	[1.0, 3.0]	[10.0, 10.0]	[0.23882922, 0.35987668]	[1.91540161, 2.29646049]
0.02	20270410	[1.0, 3.0]	[10.0, 10.0]	[0.22627741, 0.38276322]	[1.88307652, 2.37398539]

L-BFGS-B Failure D Conditions vs Matched BFGS Runs

This section compares each failing `L-BFGS-B` case against the `BFGS` run with the same noise, seed, D_{true}, and D_0. Across matched failing cases, median min logged trial-D component was 0.23 for `L-BFGS-B` vs 0.288 for `BFGS`. Median min component at last logged iterate was 1.89 for `L-BFGS-B` vs 1.01 for `BFGS`. Similarity in failures: all failed `L-BFGS-B` runs used $D_0 = [10, 10]$ and $D_{\text{true}} = [1, 3]$, with `DIVERGED_LINE_SEARCH`. Contrast: matched `BFGS` runs converged and ended at less aggressive final iterates for the same cases.

σ	Seed	D_{true}	D_0	LBFGSB min/last D	BFGS min/last D	BFGS converged
0	20270400	[1.0, 3.0]	[10.0, 10.0]	[0.2336, 0.369] / [1.893, 2.318]	[0.2847, 0.4389] / [1, 3.001]	yes
0.005	20270401	[1.0, 3.0]	[10.0, 10.0]	[0.2348, 0.3668] / [1.905, 2.32]	[0.2886, 0.44] / [1.018, 2.968]	yes
0.005	20270402	[1.0, 3.0]	[10.0, 10.0]	[0.03229, 1.014] / [1.425, 2.199]	[0.2806, 0.4436] / [0.9787, 2.927]	yes
0.005	20270404	[1.0, 3.0]	[10.0, 10.0]	[0.2357, 0.3653] / [1.916, 2.324]	[0.3034, 0.4565] / [1.029, 3.015]	yes
0.005	20270406	[1.0, 3.0]	[10.0, 10.0]	[0.04336, 1.145] / [1.418, 2.207]	[0.2873, 0.4503] / [0.9808, 3.222]	yes
0.005	20270409	[1.0, 3.0]	[10.0, 10.0]	[0.04892, 1.191] / [1.427, 2.207]	[0.3087, 0.473] / [0.9954, 2.924]	yes
0.02	20270407	[1.0, 3.0]	[10.0, 10.0]	[0.2388, 0.3599] / [1.915, 2.296]	[0.2559, 0.3826] / [1.023, 3.15]	yes
0.02	20270410	[1.0, 3.0]	[10.0, 10.0]	[0.2263, 0.3828] / [1.883, 2.374]	[0.3874, 0.6082] / [1.02, 2.606]	yes

Interpretation

- In this dataset, forward-solve failures are method-specific: observed only with `L-BFGS-B`.
- Failures cluster in the high-initial-guess regime ($D_0 = [10, 10]$), which suggests aggressive early trial steps are important.
- Failure trajectories frequently visit low-D trial values before `DIVERGED_LINE_SEARCH`; this is consistent with harder nonlinear solves in those regions.