DARTMOUTH

Exploration into Reducing Uncertainty in Inverse Problems Using Markov Chain Monte Carlo Methods

Undergraduate Thesis Presentation

Undergraduate Researcher: Jiahui Zhang ’20
Research Advisor: Dr. Anne Gelb, Department of Mathematics

Background

Inverse Problems in a Bayesian Framework

Suppose we have a **probability space** $(\Omega, \Gamma, \mathbb{P})$ and an observable random variable $\mathbf{Y} : \Omega \to \mathbb{R}^m$. Then a realization of $\mathbf{Y}$ takes the form

$$\mathbf{Y}(\omega) = y, \quad \omega \in \Omega.$$

This realization, y, is the **data**. Now suppose that random variable $\mathbf{X} : \Omega \to \mathbb{R}^n$ is the **unknown** variable of interest. Moreover, let $\mathbf{E} : \Omega \to \mathbb{R}^k$ be the **noise** random variable. For some function $g : \mathbb{R}^n \times \mathbb{R}^k \to \mathbb{R}^m$, we define the forward model as

$$\mathbf{Y} = g(\mathbf{X}, \mathbf{E}),$$

and if the noise is **additive**, we have

$$\mathbf{Y} = g(\mathbf{X}, \mathbf{E}) = g(\mathbf{X}) + \mathbf{E}.$$

Bayesian Formulation

Let $x \in \Omega$ and $e \in \Omega$ be realizations of the random variables $\mathbf{X}$ and $\mathbf{E}$ respectively. Then the **Bayes Formulation** is

$$\hat{f}_{\mathbf{X}}(x) = f_{\mathbf{X}|\mathbf{Y}}(x \,|\, y) = f_{\mathbf{Y}|\mathbf{X}}(y \,|\, x)\tilde{f}_{\mathbf{X}}(x),$$

where $\hat{f}$ is the **posterior** probability density function, $f_{\mathbf{Y}|\mathbf{X}}(y \,|\, x)$ is the **likelihood** probability density function, and $\tilde{f}_{\mathbf{X}}(x)$ is the **prior** probability density function.

The goal is to develop methods to explore the **posterior probability density** $\hat{f}_{\mathbf{X}}(x) = f_{\mathbf{X}|\mathbf{Y}}(x \,|\, y)$ so to obtain **single estimates** or **distributions**.

ℓ_2 and ℓ_1 Prior

For the ℓ_2 prior, the probability density function is defined as

$$\tilde{f}_{\mathbf{X}}(x) = (\frac{1}{2\pi \text{det}(\Sigma)})^{\frac{n}{2}} \exp\Big(-\frac{1}{2}(x - x_0)^T \Sigma^{-1} (x - x_0)\Big),$$

for unknown $\mathbf{X}$ with mean $x_0 \in \mathbb{R}^n$ and covariance matrix $\Sigma \in \mathbb{R}^{n \times n}$.

For the ℓ_1 prior, the probability density function is defined as

$$\tilde{f}_{\mathbf{X}}(x) = \left(\frac{\alpha}{2}\right)^2 \exp(-\alpha ||x||_1),$$

for some constant $\alpha > 0$.

Additive Noise

Suppose that the variance of $\mathbf{E}$ is σ and the forward model is **linear**. Then $g(\mathbf{X}) = A\mathbf{X}$ for some matrix $A \in \mathbb{R}^{m\times n}$ and the forward model is

$$\mathbf{Y} = g(\mathbf{X}) + \mathbf{E} = A\mathbf{X} + \mathbf{E},$$

and the distribution of the error obeys the following density function

$$f_{\mathbf{E}}(e) = C_1 \exp\left\{ -\frac{||e||_2^2}{2\sigma^2} \right\},$$

for some $C_1 \in \mathbb{R}$.

Posterior Probability Density

Assuming independence of the noise and unknown, the **likelihood function** is

$$f_{\mathbf{Y}|\mathbf{X}}(y\,|\,x) = f_{\mathbf{E}}(y - Ax) = C_1 \exp\left\{ -\frac{||y - Ax||_2^2}{2\sigma^2} \right\}.$$

By Bayes Theorem, we have the **posterior probability density function**

$$\hat{f}_{\mathbf{X}}(x) = C_3 \exp\left\{-\alpha\,||x||_1 - \frac{1}{2\sigma^2}||y - Ax||_2^2\right\},$$

for some constant $C_3 \in \mathbb{R}$.

The parameters α and σ corresponds to the noise present and are chosen to be 1. Tuning these parameters become unnecessary when we employ the technique of **Variance Based Joint Sparsity (VBJS)**.

Variance Based Joint Sparsity (VBJS)

- We assume the unknown function is **piecewise smooth** and employ the **polynomial annihilation operator** in the **prior** to transform the unknown into the **sparse edge domain (this is useful in practice when considering imaging applications)**.

- **VBJS (Gelb, Scarnati)** provides a way to spatially vary parameters that penalize smooth regions more than regions with edges.

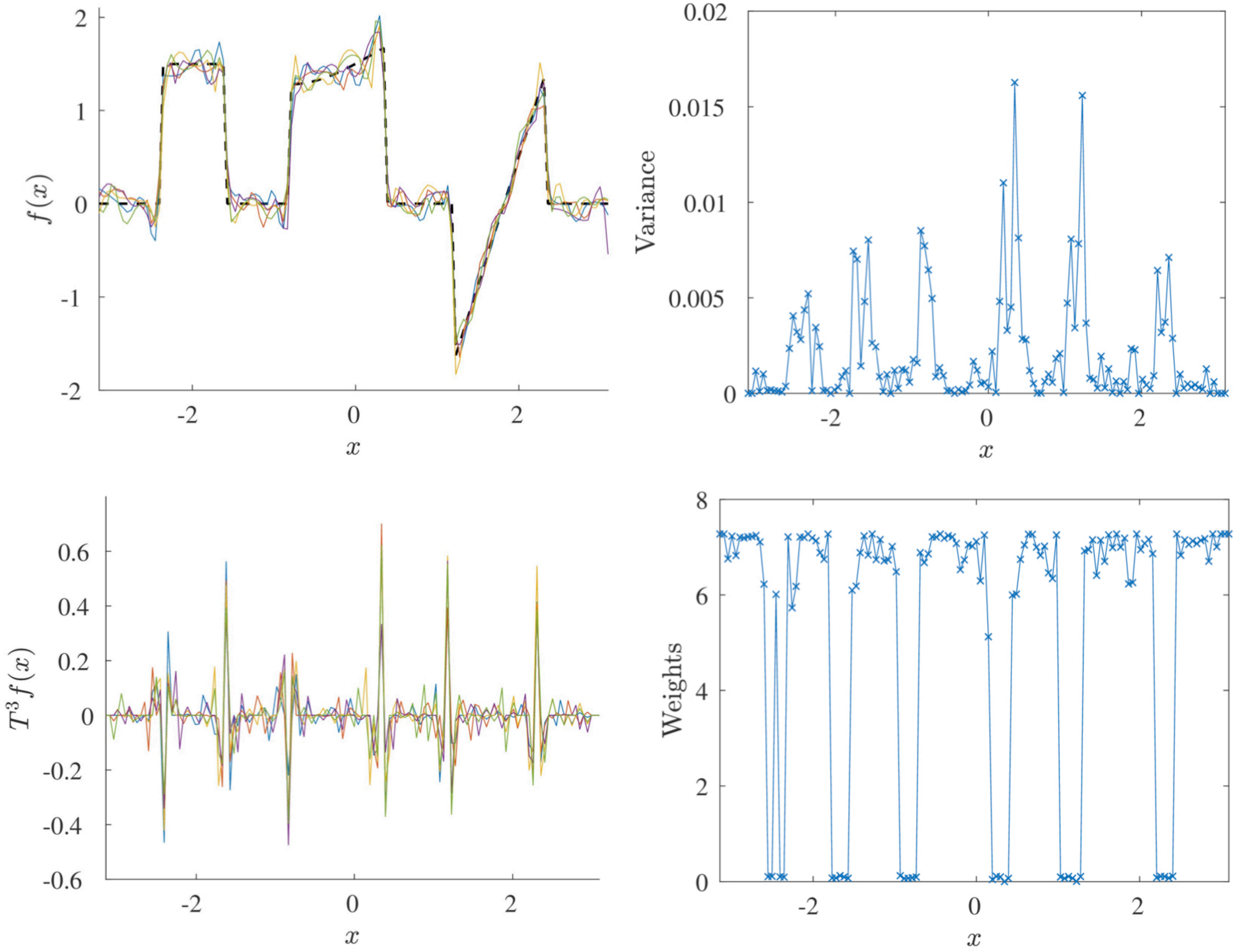
- The method tests the variance of **multiple measurement vectors** at different points of the domain.

- A large variance in a region of the **edge domain** may suggest an edge in that region.

- This variance information is used to construct **weights** for the prior in which greater weights are placed in regions of smaller variance.

- The weights make tuning the parameters α and σ unnecessary.

Variance Based Joint Sparsity (VBJS)



- The top-left figure shows the set of **multiple measurement vectors** and the true unknown function.

- The bottom-left figure shows the multiple measurement vectors in their **sparse domain** (edge domain).

- The top-right figure shows the **variances** of he multiple measurement vectors in the sparse domain at each region.

- The bottom-right shows the **normalized weights** constructed through the variances.

Variance Based Joint Sparsity (VBJS)

By imposing the prior belief that any realization x of $\mathbf{X}$ should be sparse in the sparse domain $\mathscr{L}x$, we can modify the the posterior distribution to be

$$\hat{f}_{\mathbf{X},w}(x) = C_3 \exp\left\{-\alpha ||W\mathscr{L}x||_1 - \frac{1}{2\sigma^2}||y - Ax||_2^2\right\},$$

where $W = diag(w)$ is a fixed matrix and $\mathscr{L}$ is the **sparsifying operator**.

We chose the ℓ_1 prior for the posterior density function because it enforces sparsity in the edge domain.

Now, we can obtain a point estimate with a **maximum *a priori* (MAP) estimation**

$$x_{MAP} = \text{argmin}_{x\in\mathbb{R}^n}\left\{\alpha ||Lx||_1 + \frac{1}{2\sigma^2}||y - Ax||_2^2\right\}.$$

Markov Chain Monte Carlo (MCMC)

MCMC is the statistical technique used to recover the posterior probability density $\hat{f}_{\mathbf{X}}(x)$.

First a discrete Markov chain is a **stochastic process** where given the present state, past and future states are independent. This property is said to be the **Markov property** and can be stated as

$$\mathbb{P}(\theta^{(n+1)} \in A \ | \theta^{(n)} = x, \theta^{(n-1)} \in A_{n-1}, \dots, \theta^{(0)} \in A_0) = \mathbb{P}(\theta^{(n+1)} \in A \,|\, \theta^{(n)} = x),$$

for all sets $A_0, A_1, \dots, A_{n-1}, A \in \Omega$ and $x \in \Omega$.

Further, for a finite Markov Chain $(\theta)^t$ where the sample space is $\Omega = \{x_1, x_2, x_3, \dots, x_n\}$, there exists a **kernel** $p(i, j)$ represented by some matrix $K \in \mathbb{R}^{n \times n}$ such that

$$\mathbb{P}(\theta^{n+1} = x_j \,|\, \theta^n = x_i) = p(i, j) = [K]_{i,j}.$$

Further, we define a special type of Markov chain that satisfies **ergodicity,** which means it is **irreducible** and **aperiodic**.

Ergodicity ensures that the Markov chain possesses a **unique stationary distribution**.

Markov Chain Monte Carlo (MCMC)

Another property we must introduce is the **detailed balance condition**. Suppose that the transition kernel of an irreducible Markov chain is defined to be $p(\,\cdot\mid\cdot\,)$ then the Markov chain has *detailed balance* if for some distribution $\pi \in \Omega$

$$\pi(\theta)p(\theta, \phi) = \pi(\phi)p(\phi, \theta).$$

This is a much stronger assumption than global balance and any distribution π that satisfies detailed balance is also the stationary distribution. Therefore, if $\overrightarrow{x_s}$ is the vector representation of the stationary distribution π,

$$\lim_{t\to\infty} K^t \overrightarrow{x} = \overrightarrow{x_s}.$$

Metropolis-Hastings Algorithm

Suppose we wish to recover the distribution π. We build a Markov chain $(X)^t$ as follows. First there must be a **proposal distribution** $q(\,\cdot\,,\cdot\,) : \Omega \times \Omega \to [0,1]$ such that $q(x^{cand} \,|\, x^{n-1})$ gives the probability the state x^{cand} is proposed given that the current state is x^{n-1}.

Now, consider x^{cand} the candidate state proposed by x^{n-1}. We accept this candidate to be the next state x^n with the **acceptance probability**

$$\alpha(x^{cand} \,|\, x^{n-1}) = \min\left\{1, \quad \frac{q(x^{n-1} \,|\, x^{cand})\pi(x^{cand})}{q(x^{cand} \,|\, x^{n-1})\pi(x^{n-1})}\right\}.$$

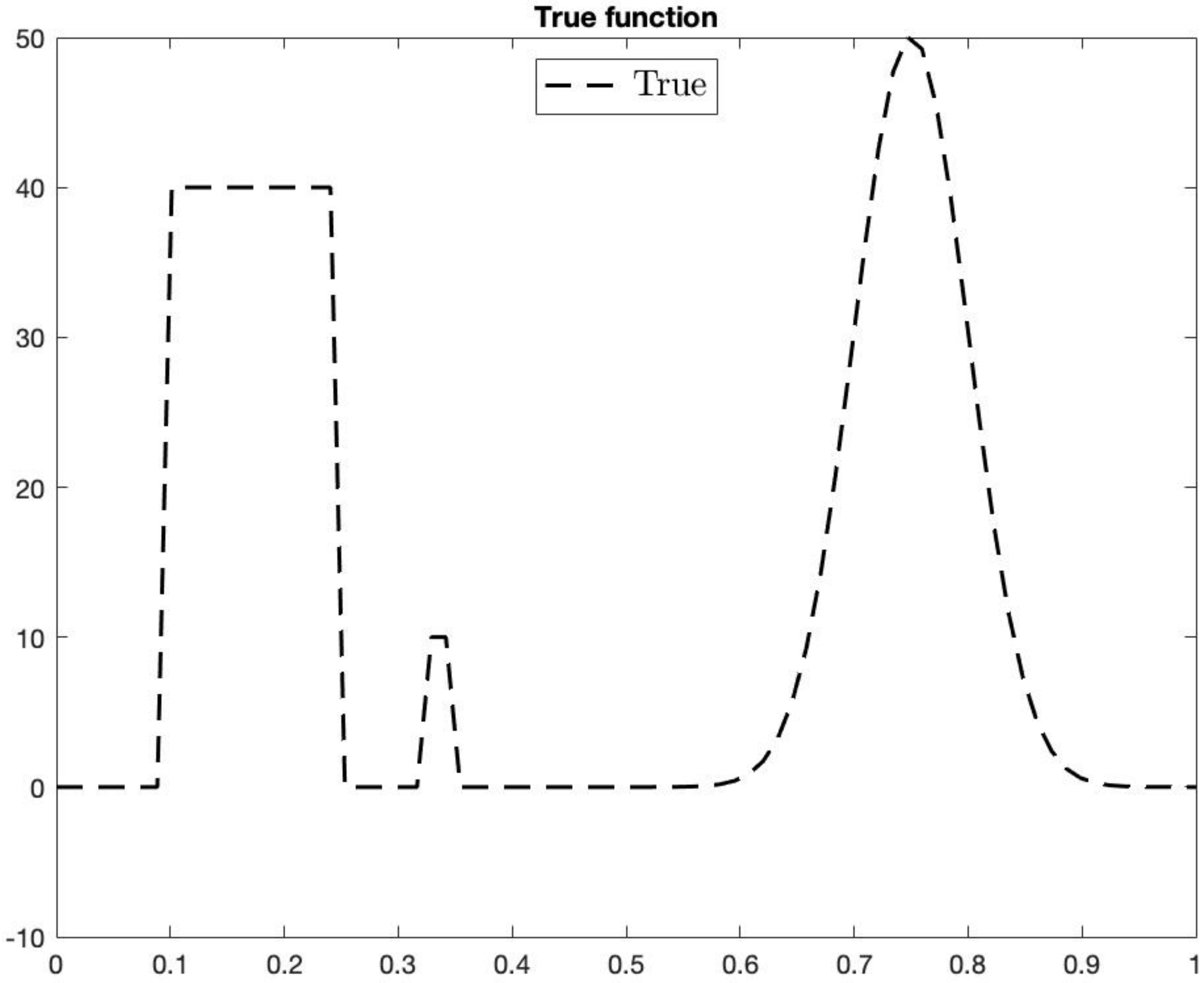
Now, the transition probability of this Markov chain is defined as

$$p(x_i, x_j) = q(x_i, x_j)\alpha(x_i, x_j),$$

where $i \neq j$.

Methodology

One-dimensional Problem



The unknown function to be recovered

The one-dimensional problem investigated in this thesis is of the form

$$\mathbf{Y} = A\mathbf{X} + \mathbf{E},$$

in which the domain is $\Omega = \mathbb{R}^{80}$.

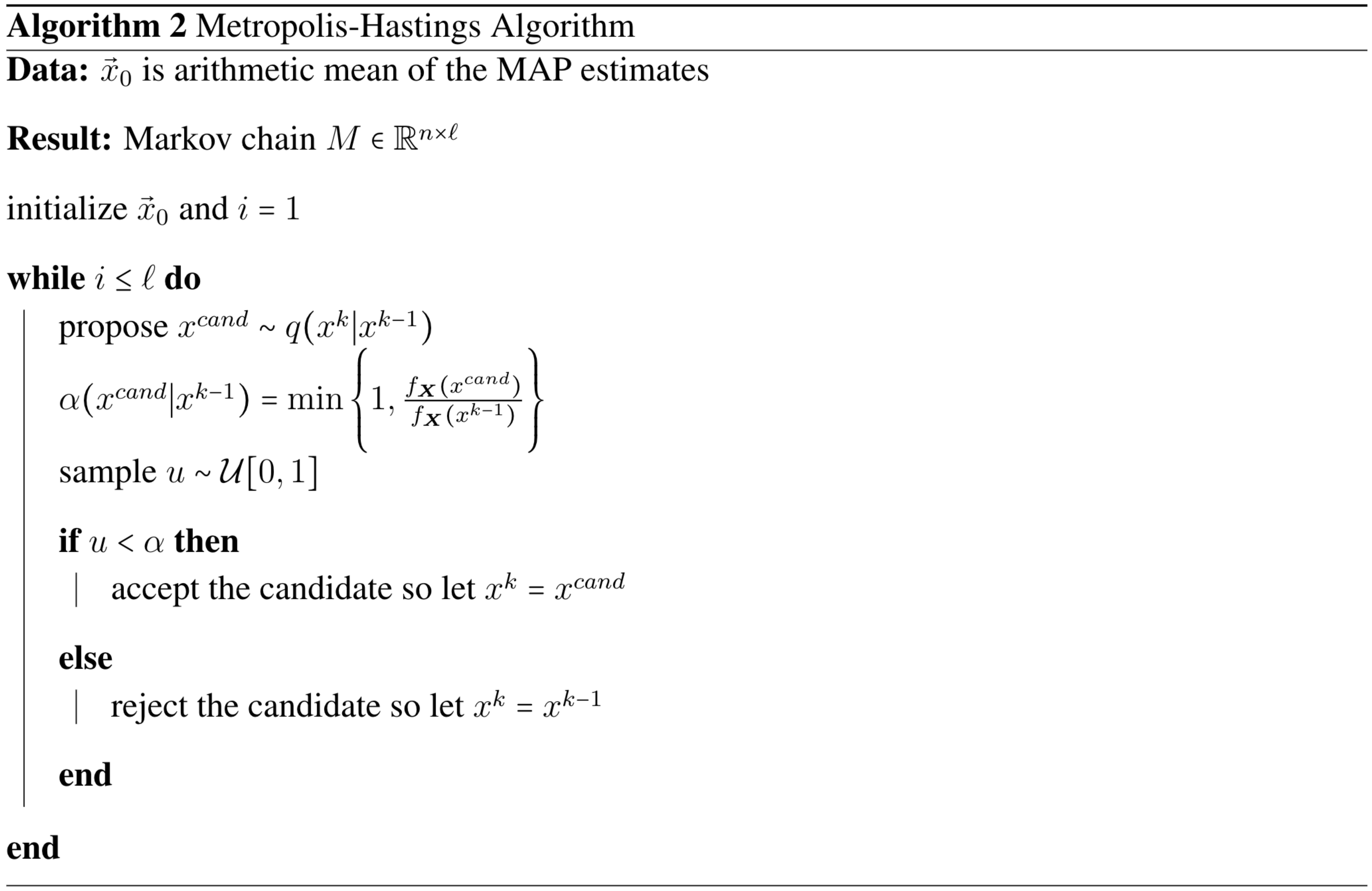
$\mathbf{X} : \Omega \to \mathbb{R}$ is the unknown random variable,
$\mathbf{Y} : \Omega \to \mathbb{R}$ is the observable random variable, and
$\mathbf{E} : \Omega \to \mathbb{R}$ is the Gaussian noise.

The forward operator $A \in \mathbb{R}^{80\times 80}$ is of the form

$$[A]_{ij} = \frac{1}{n}\frac{\exp(\frac{-(\frac{i-j}{n})^2}{2\gamma^2})}{\sqrt{\pi\gamma^2}}.$$

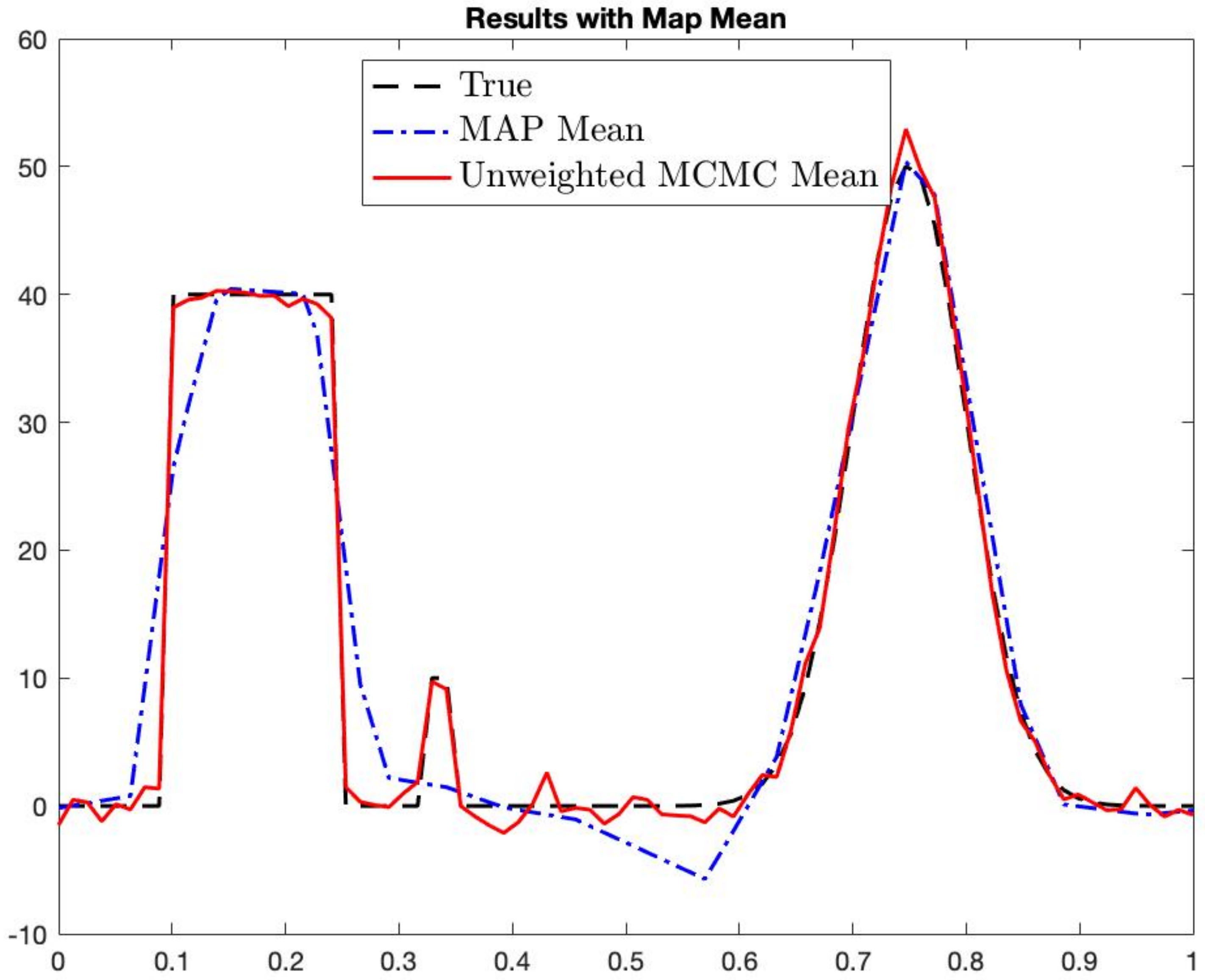
We sample from $\mathbf{Y}$ to obtain 20 measurements and obtain their respective MAP estimates.

Metropolis-Hastings Algorithm

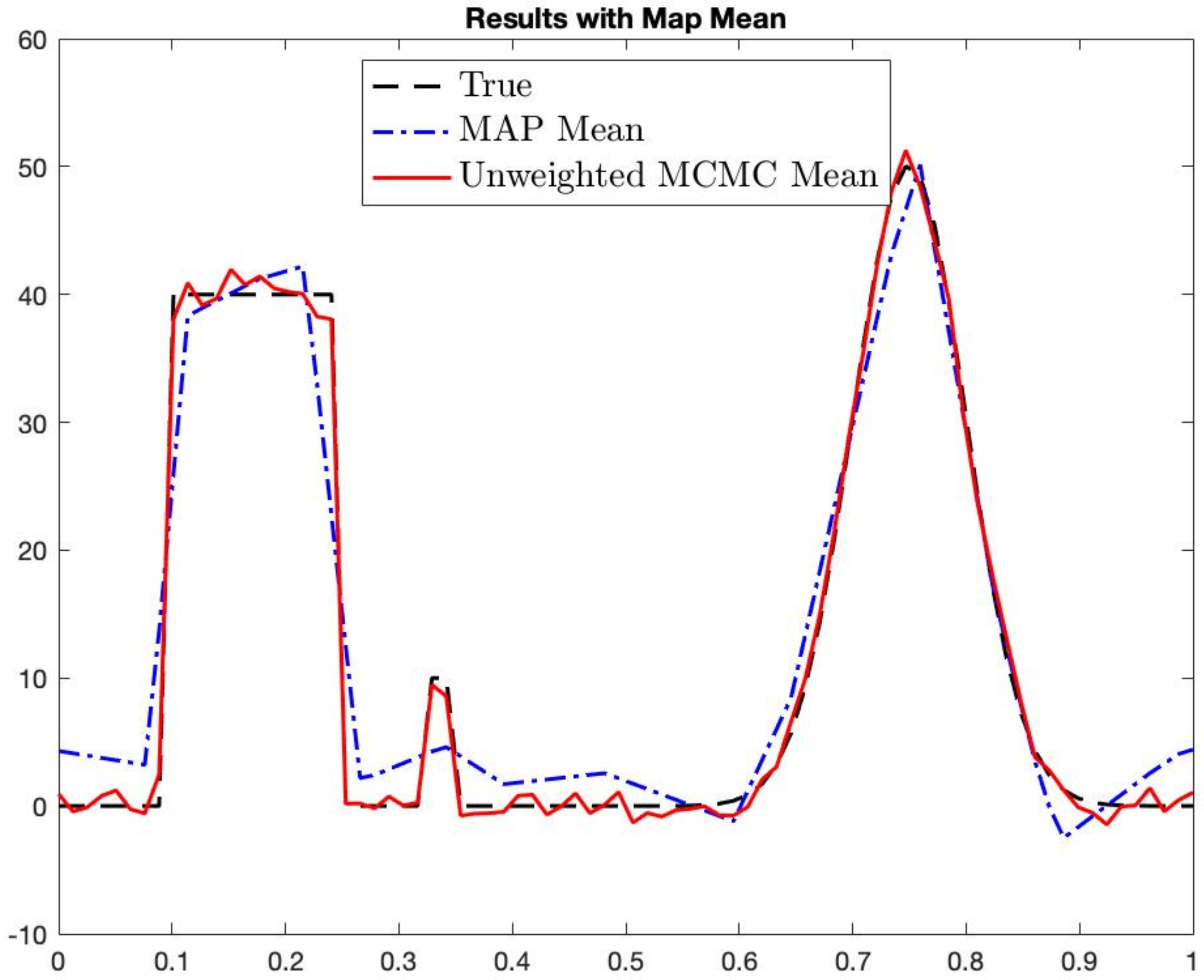


Algorithm 2 Metropolis-Hastings Algorithm

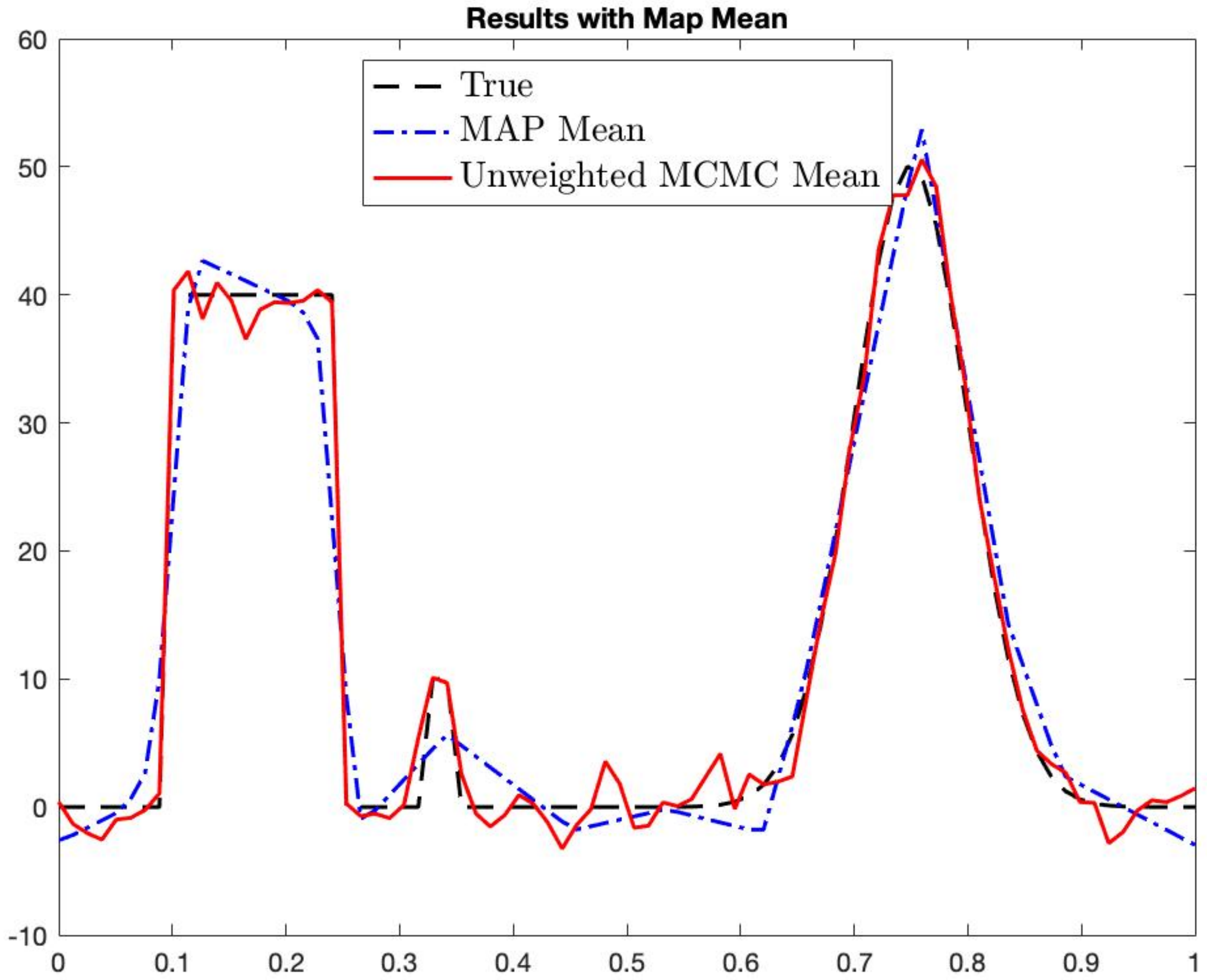
```
Data: $\vec{x}_0$ is arithmetic mean of the MAP estimates
Result: Markov chain $M \in \mathbb{R}^{n \times \ell}$
initialize $\vec{x}_0$ and $i = 1$
while $i \leq \ell$ do
    propose $x^{cand} \sim q(x^k | x^{k-1})$
    $\alpha(x^{cand} | x^{k-1}) = \min\left\{1, \frac{f_{\boldsymbol{X}}(x^{cand})}{f_{\boldsymbol{X}}(x^{k-1})}\right\}$
    sample $u \sim \mathcal{U}[0, 1]$
    if $u < \alpha$ then
        accept the candidate so let $x^k = x^{cand}$
    else
        reject the candidate so let $x^k = x^{k-1}$
    end
end
```



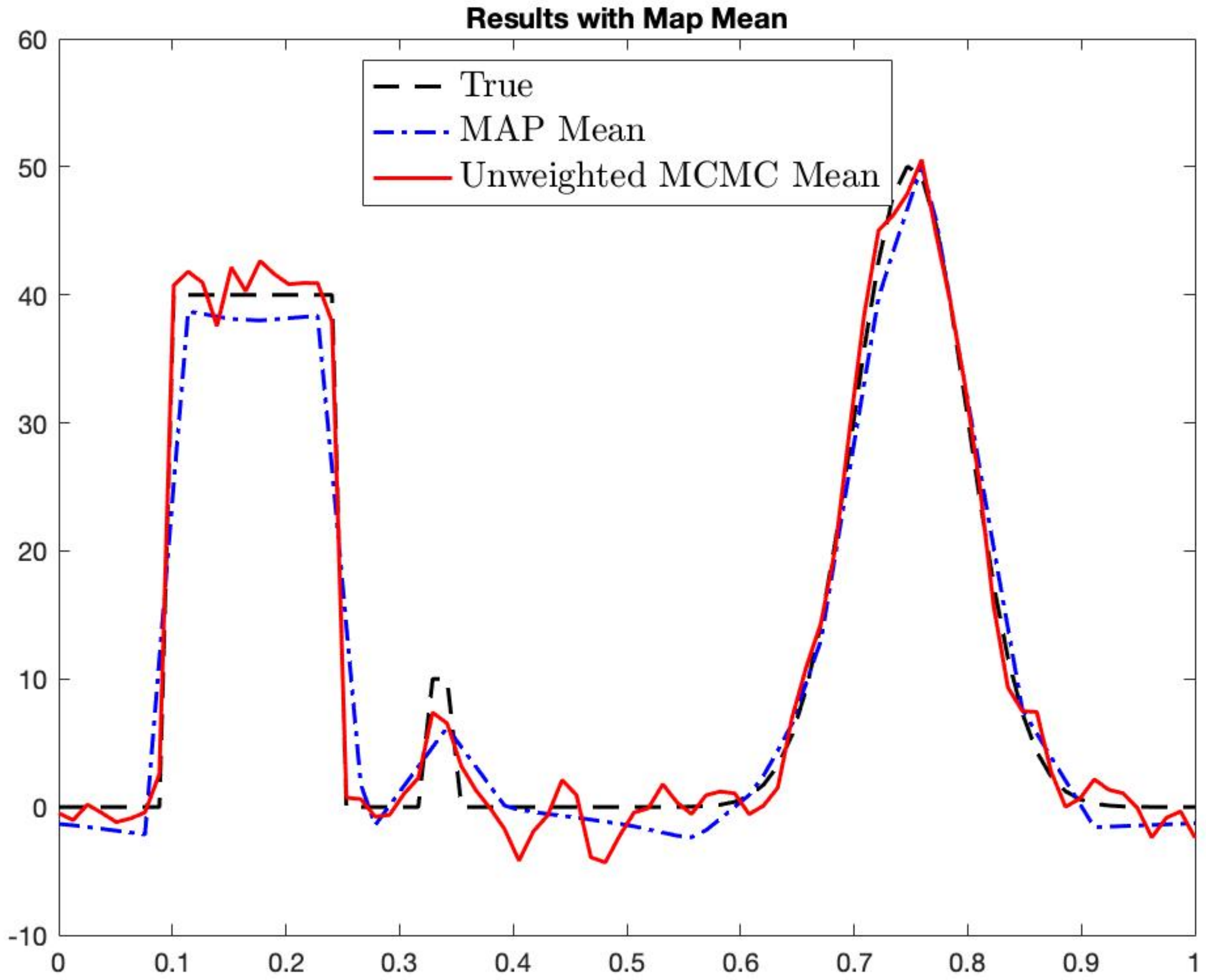
$\sigma = 0.25/\mathsf{SNR} = 36.75$



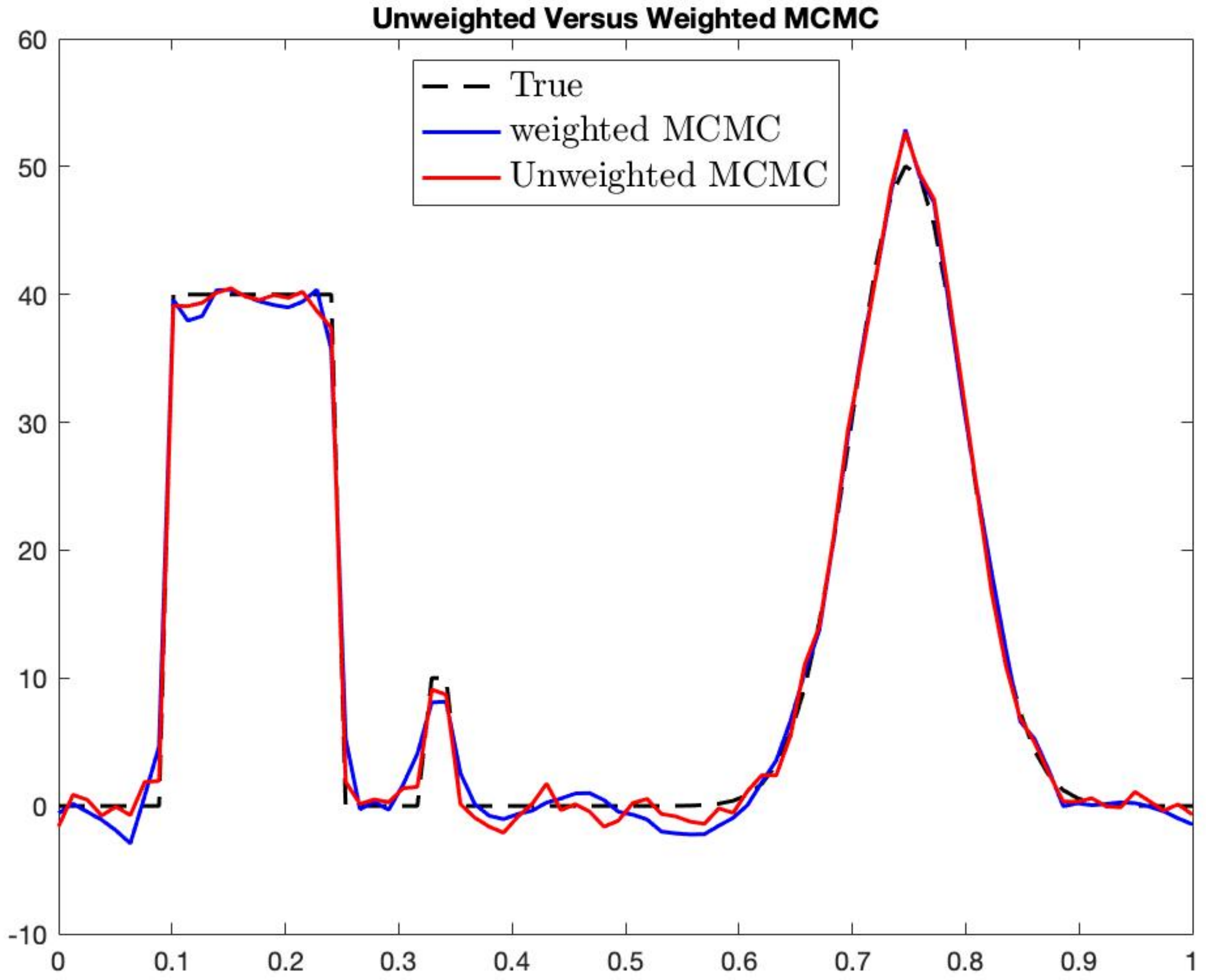
$\sigma = 0.50/\mathsf{SNR} = 32.20$



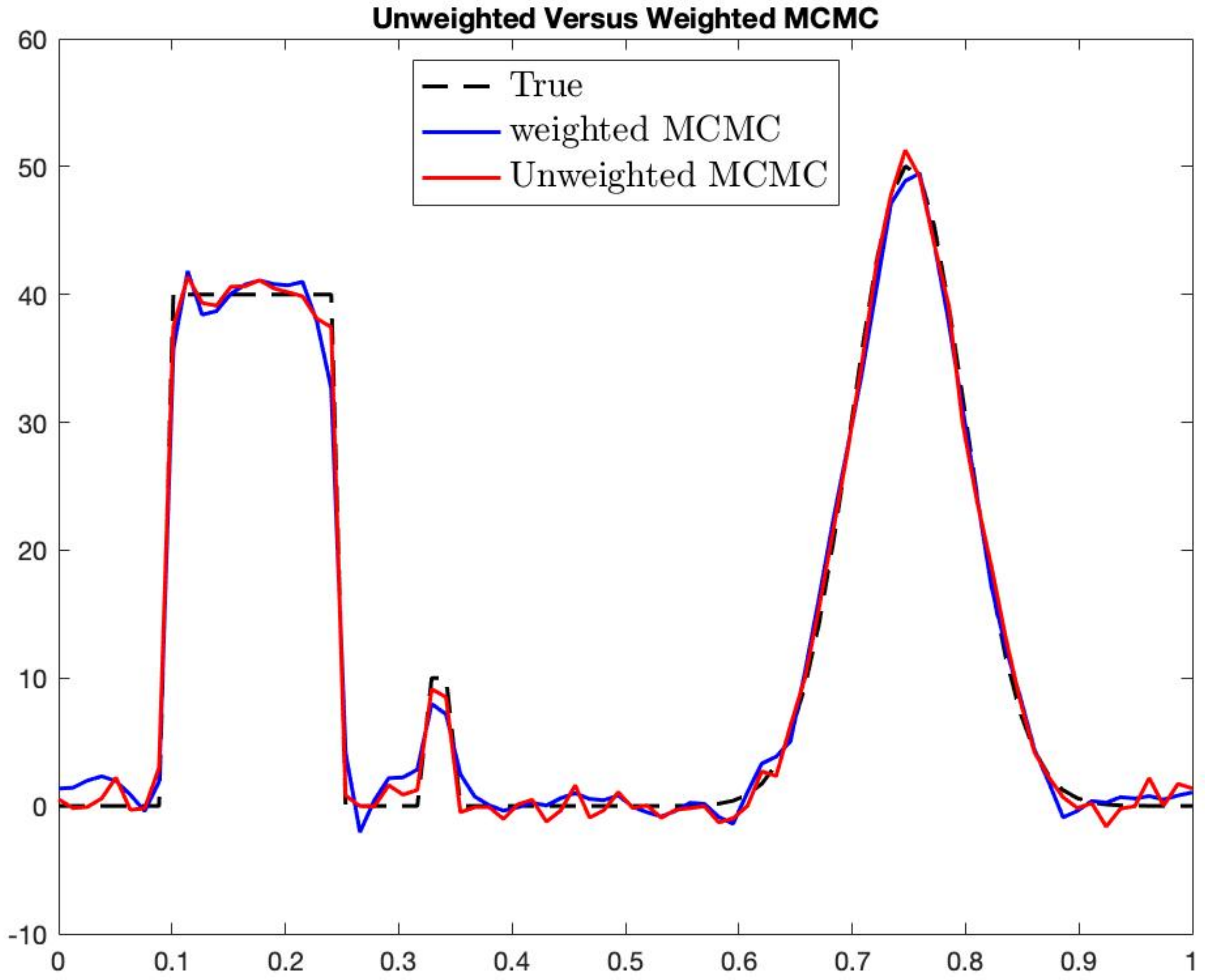
$\sigma = 0.75/\mathsf{SNR} = 27.43$



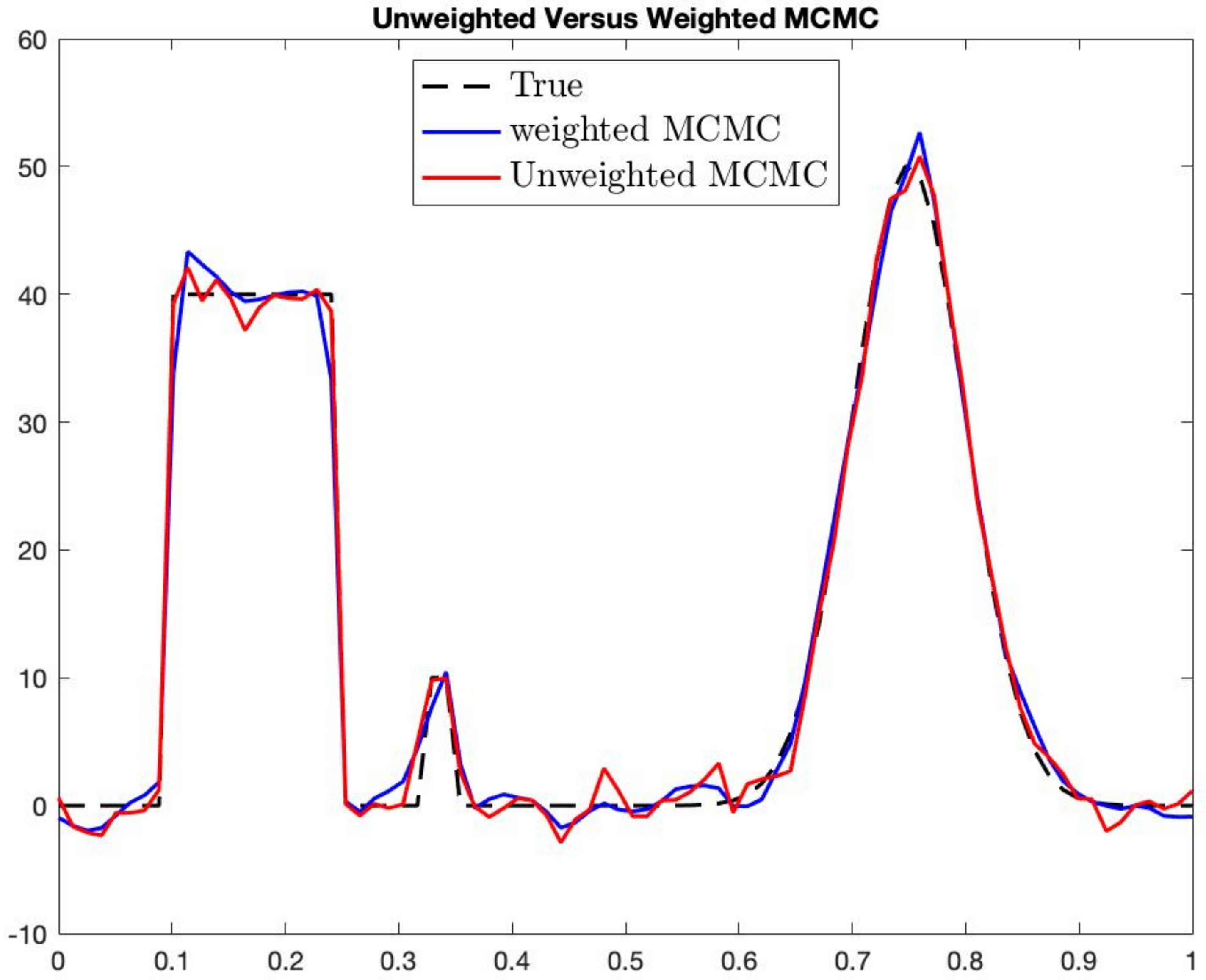
$\sigma = 1.00/\mathsf{SNR} = 25.13$



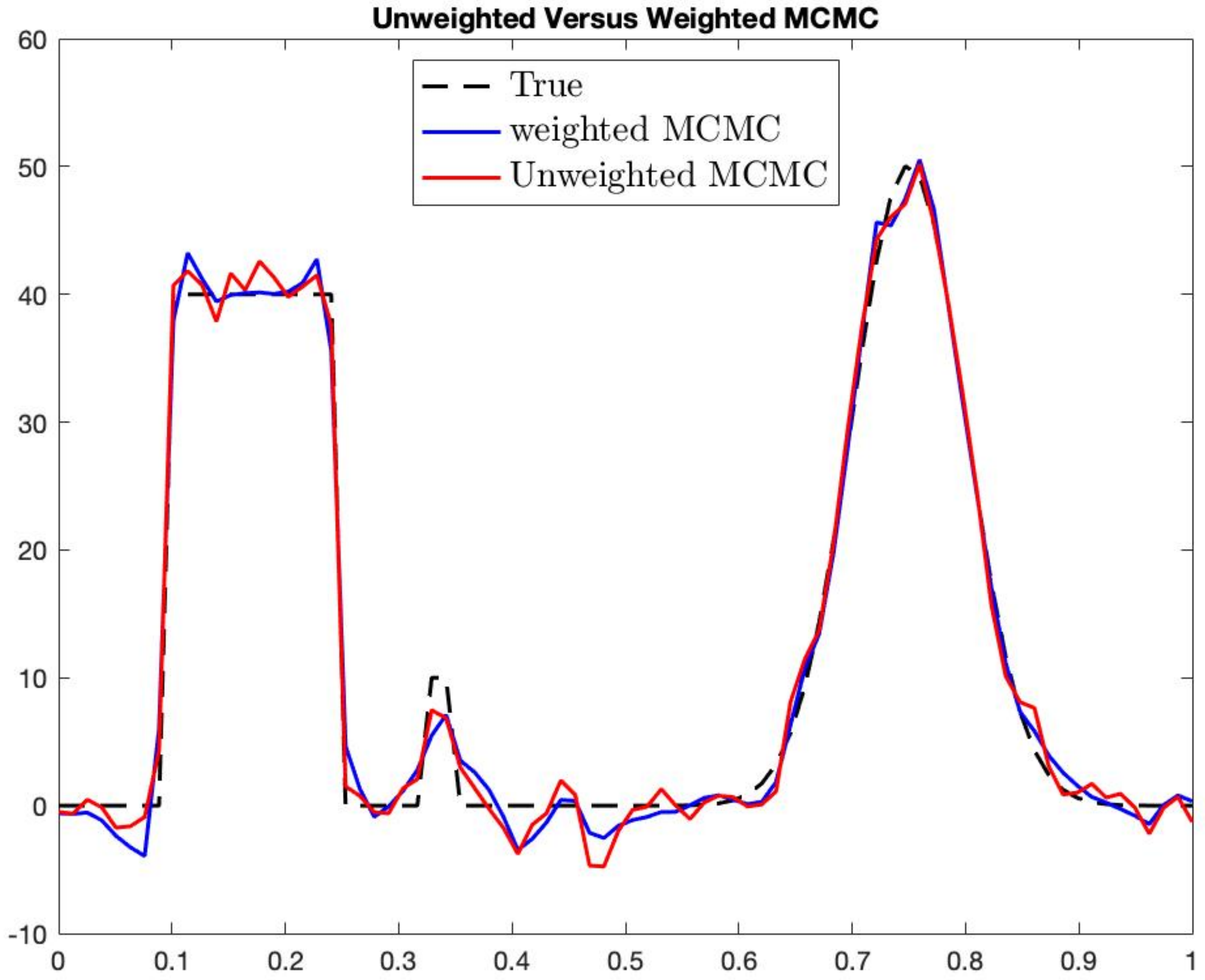
$\sigma = 0.25/\mathsf{SNR} = 36.75$



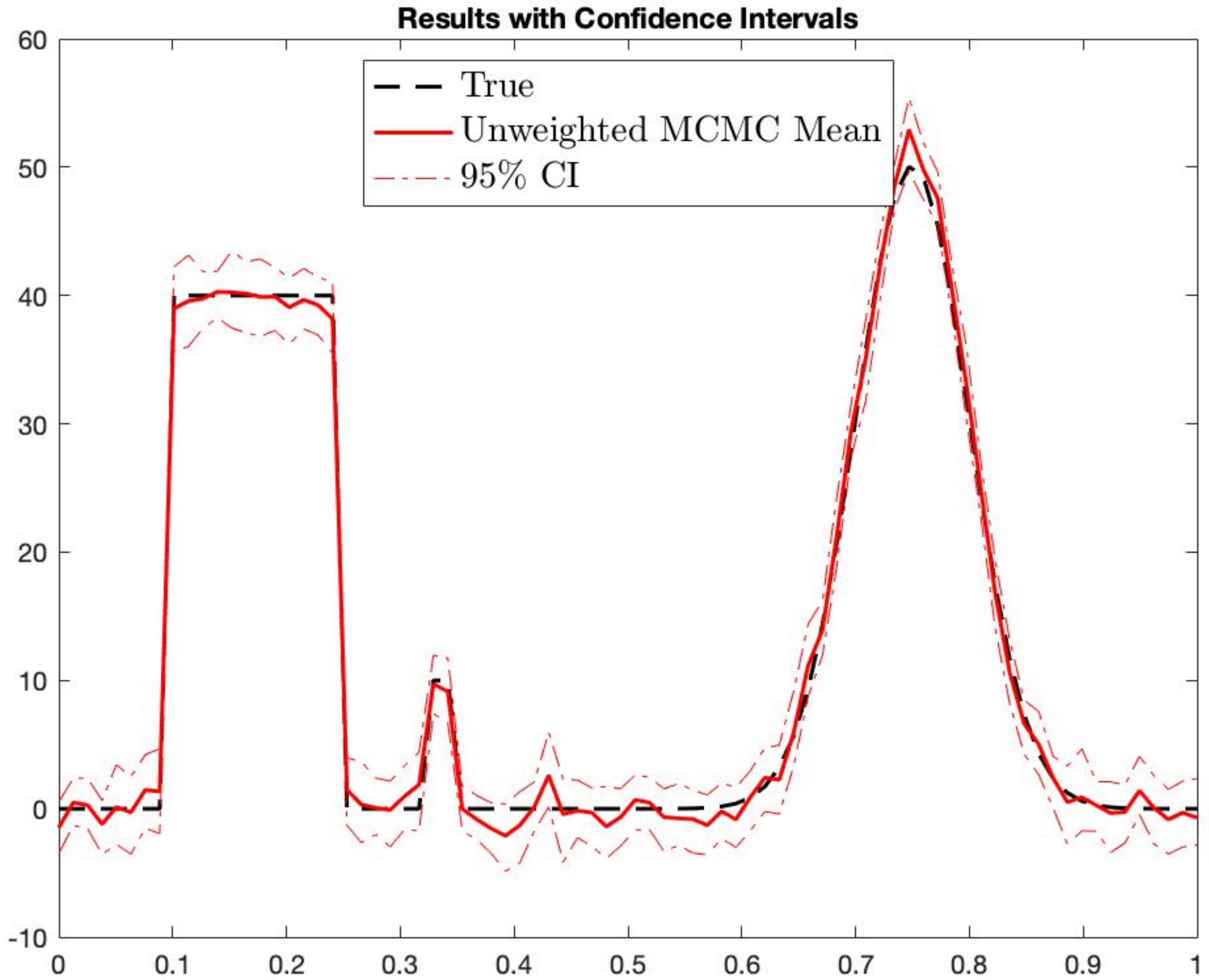
$\sigma = 0.50/\mathsf{SNR} = 32.20$



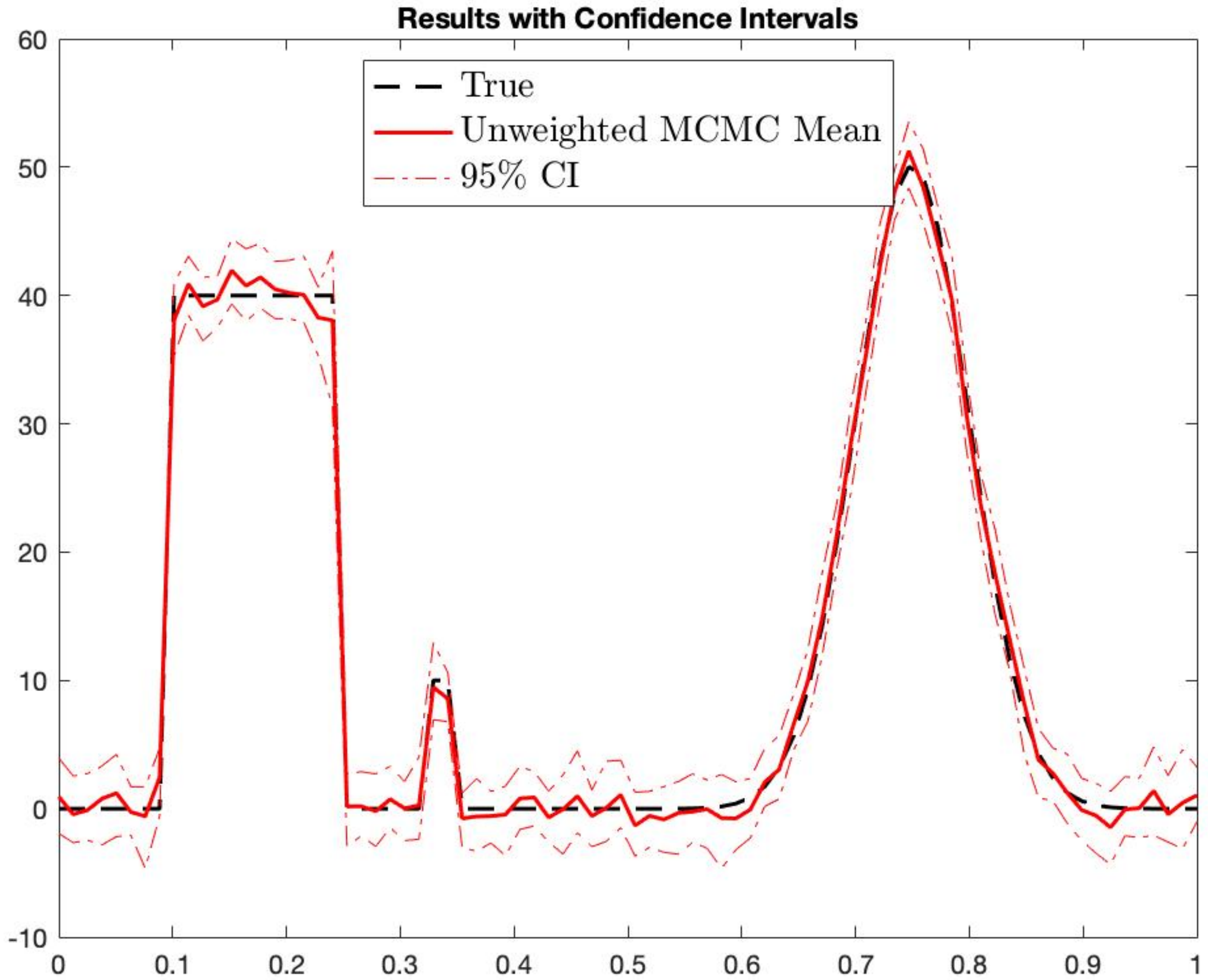
$\sigma = 0.75/\mathsf{SNR} = 27.43$



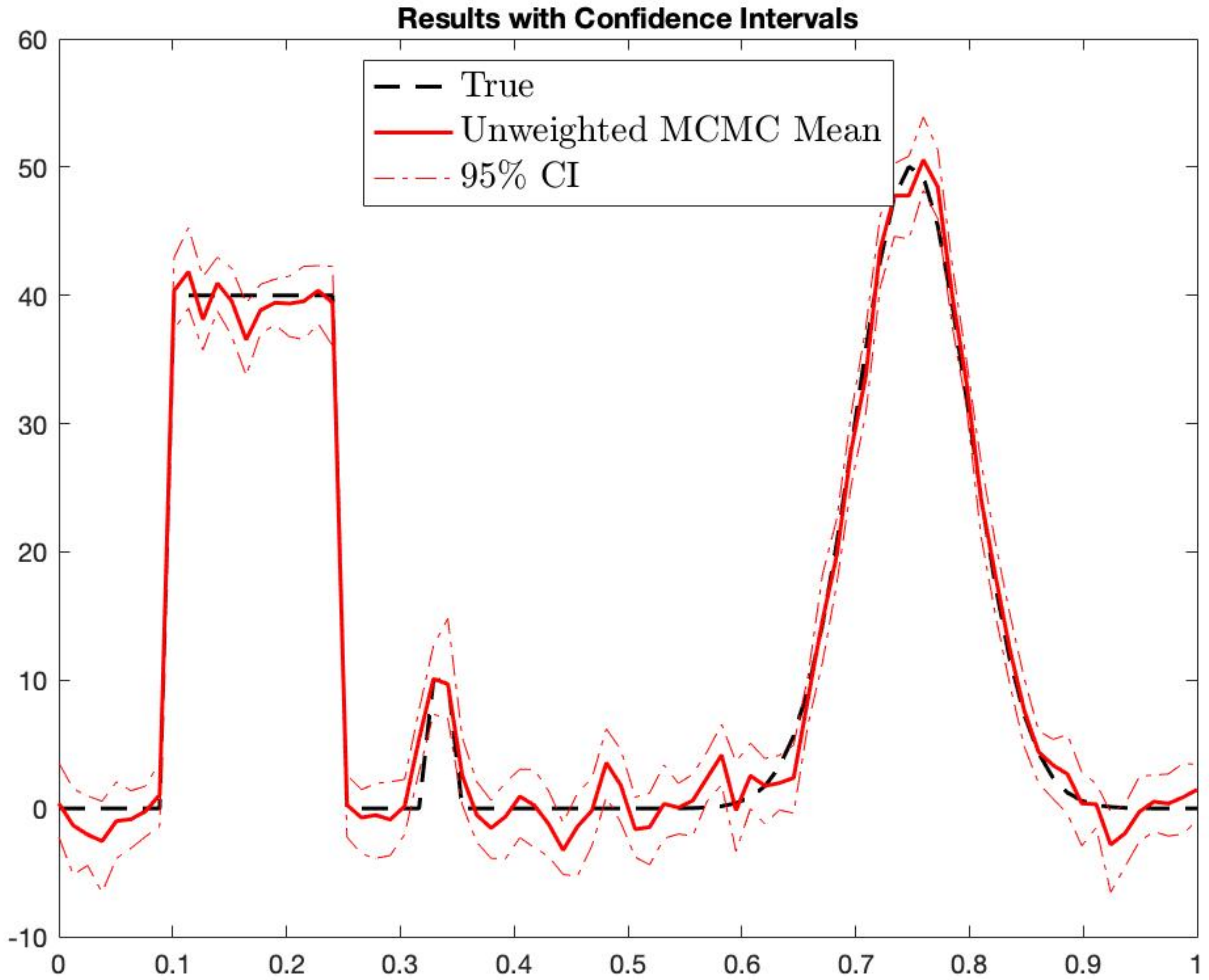
$\sigma = 1.00/\mathsf{SNR} = 25.13$



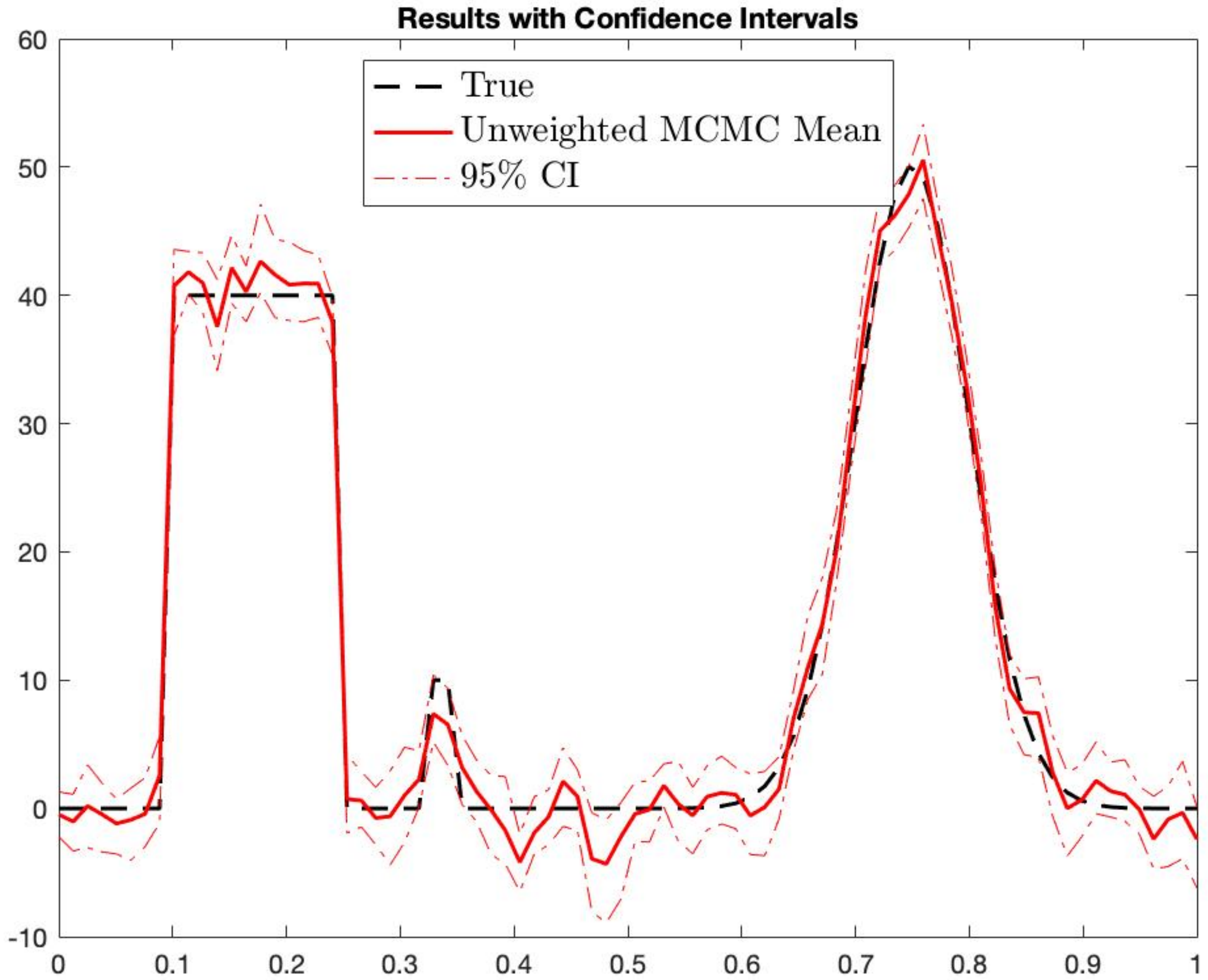
$\sigma = 0.25/\mathsf{SNR} = 36.75$



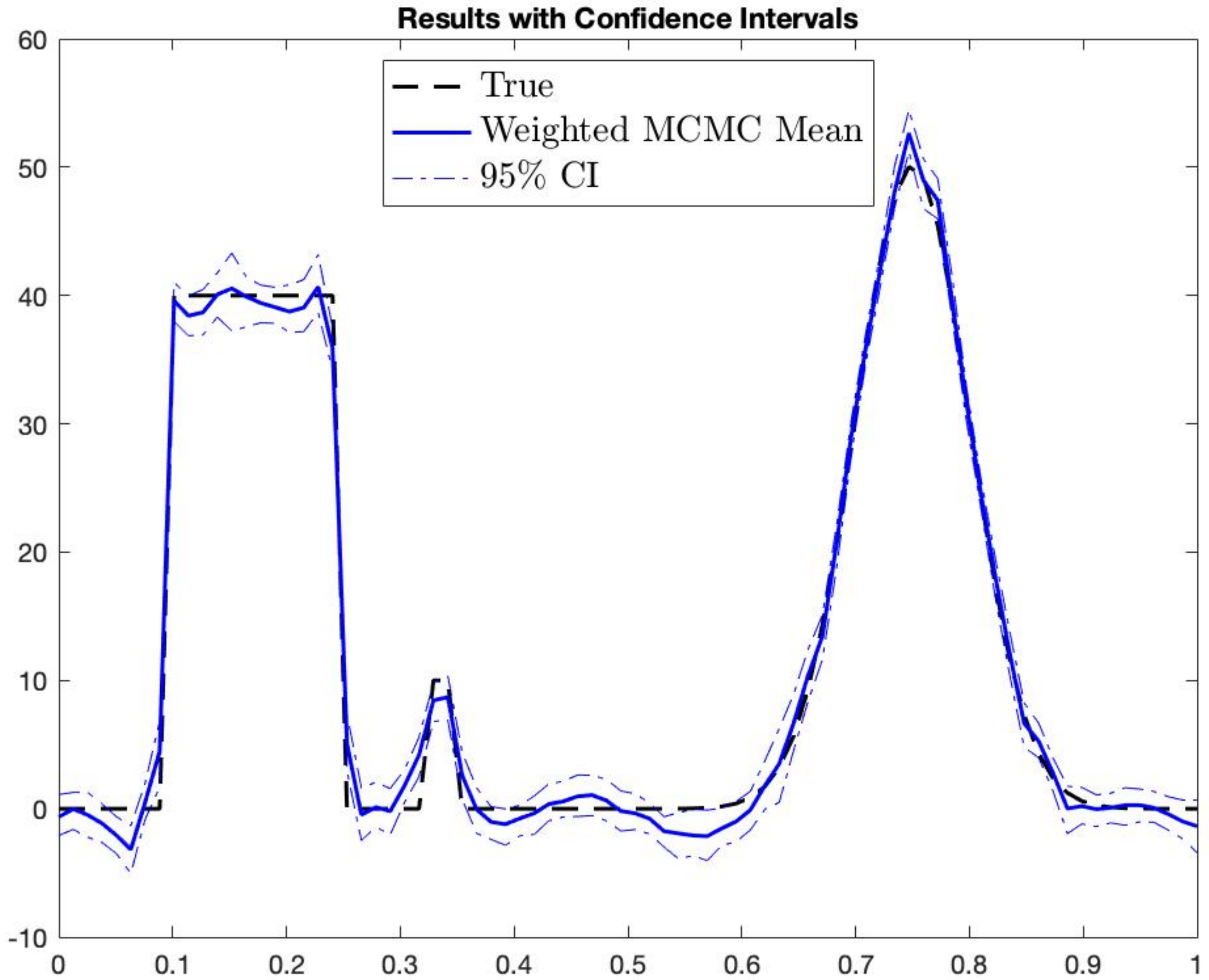
$\sigma = 0.50/\mathsf{SNR} = 32.20$



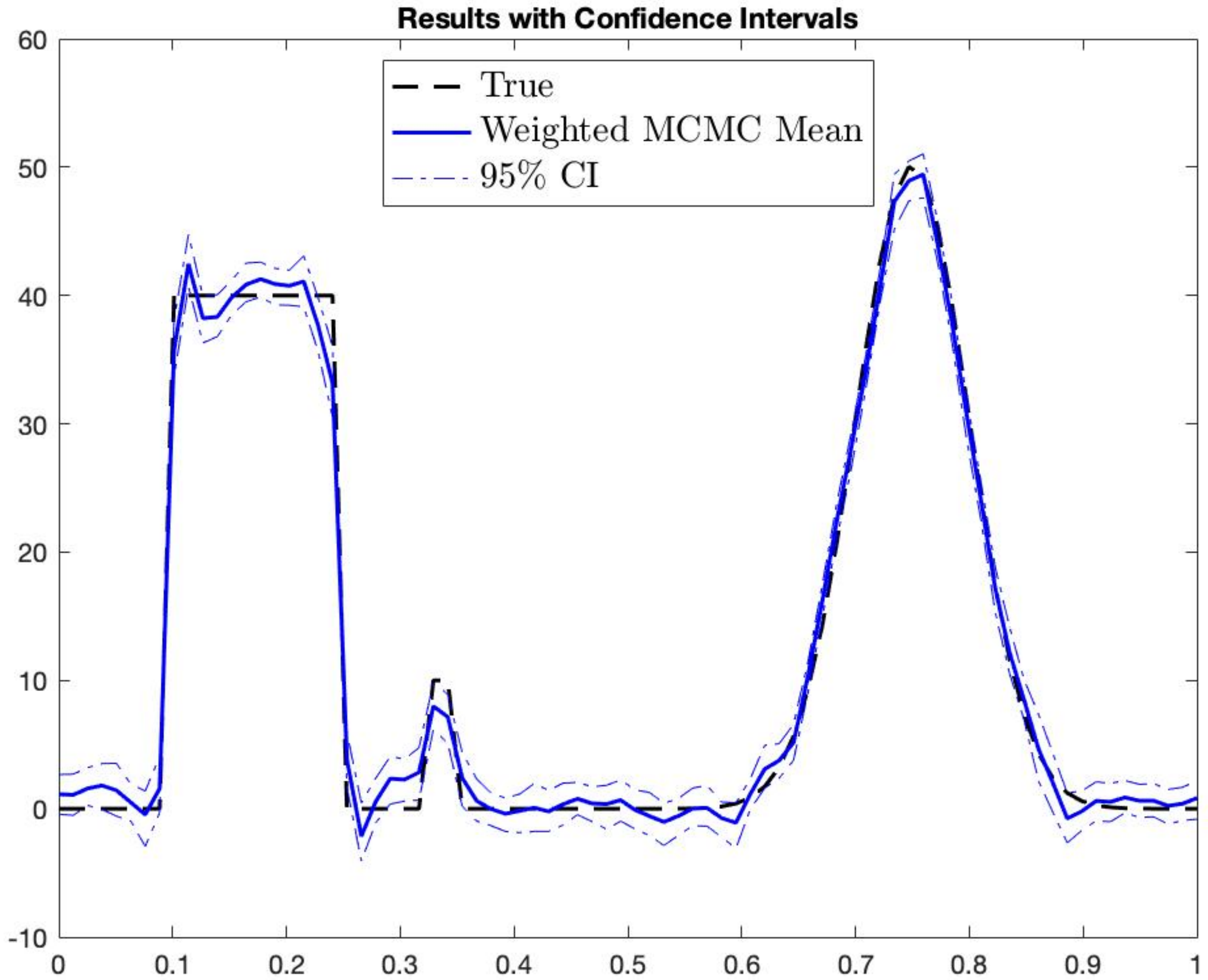
$\sigma = 0.75/\mathsf{SNR} = 27.43$



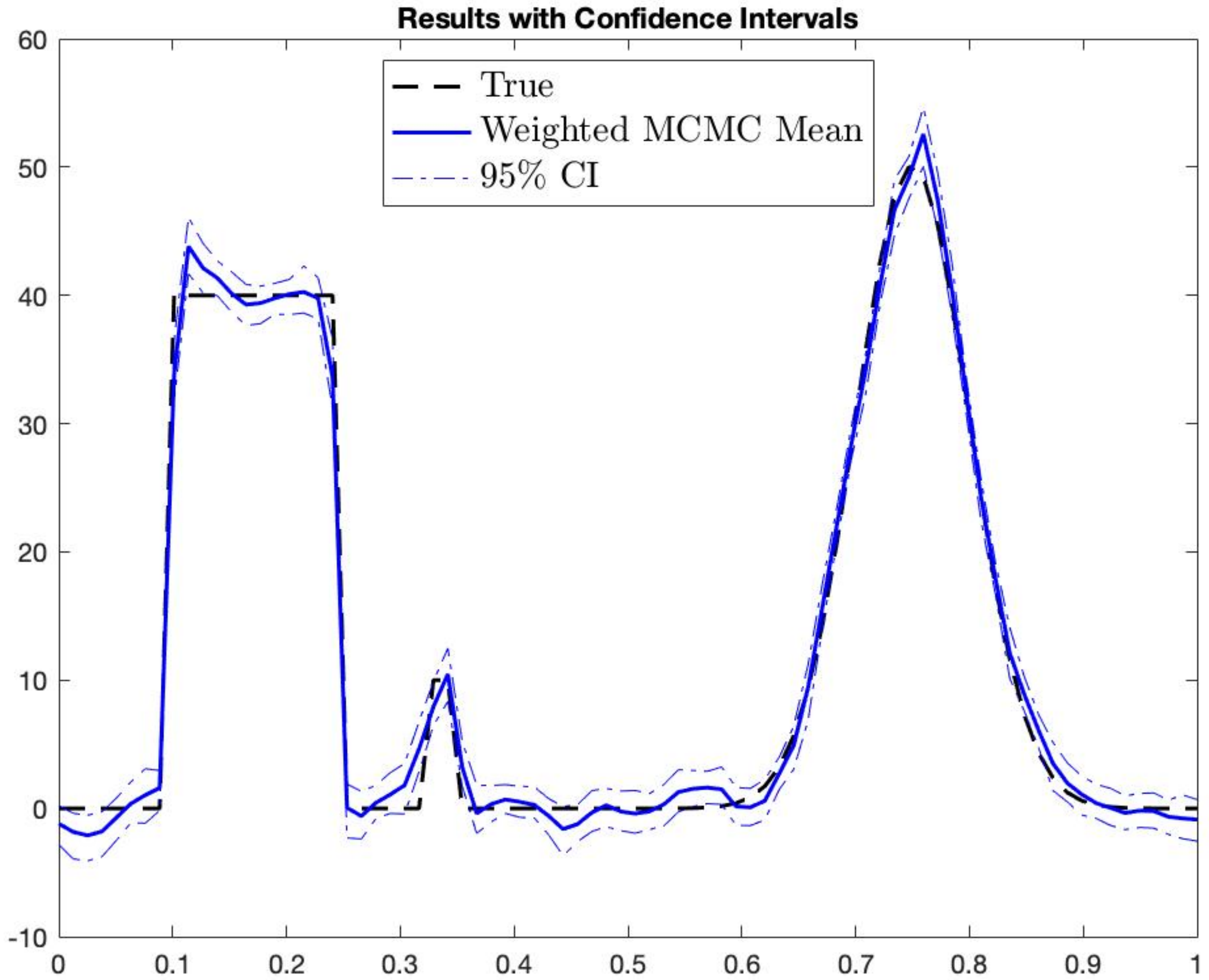
$\sigma = 1.00/\mathsf{SNR} = 25.13$



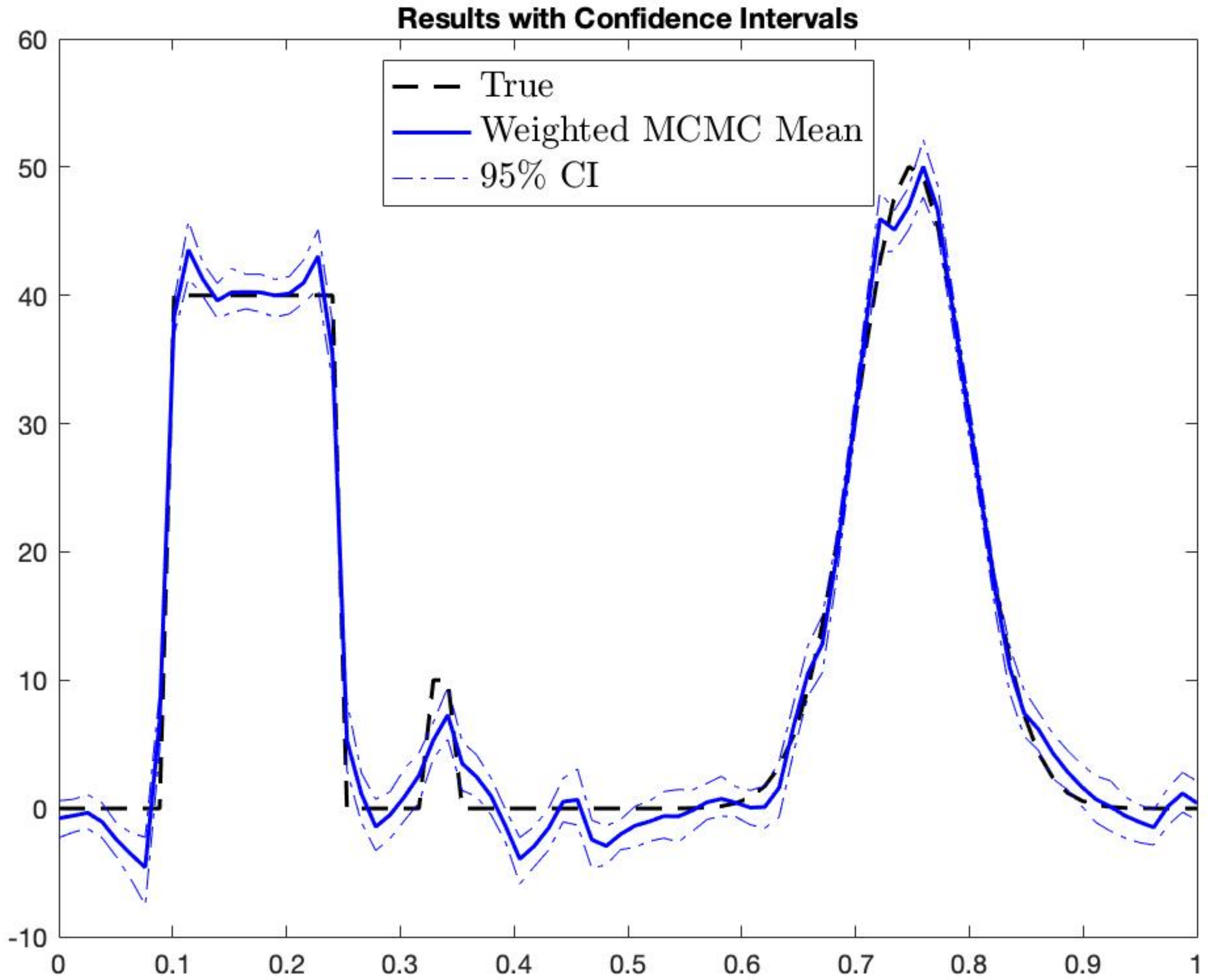
$\sigma = 0.25/\mathsf{SNR} = 36.75$



$\sigma = 0.50/\mathsf{SNR} = 32.20$

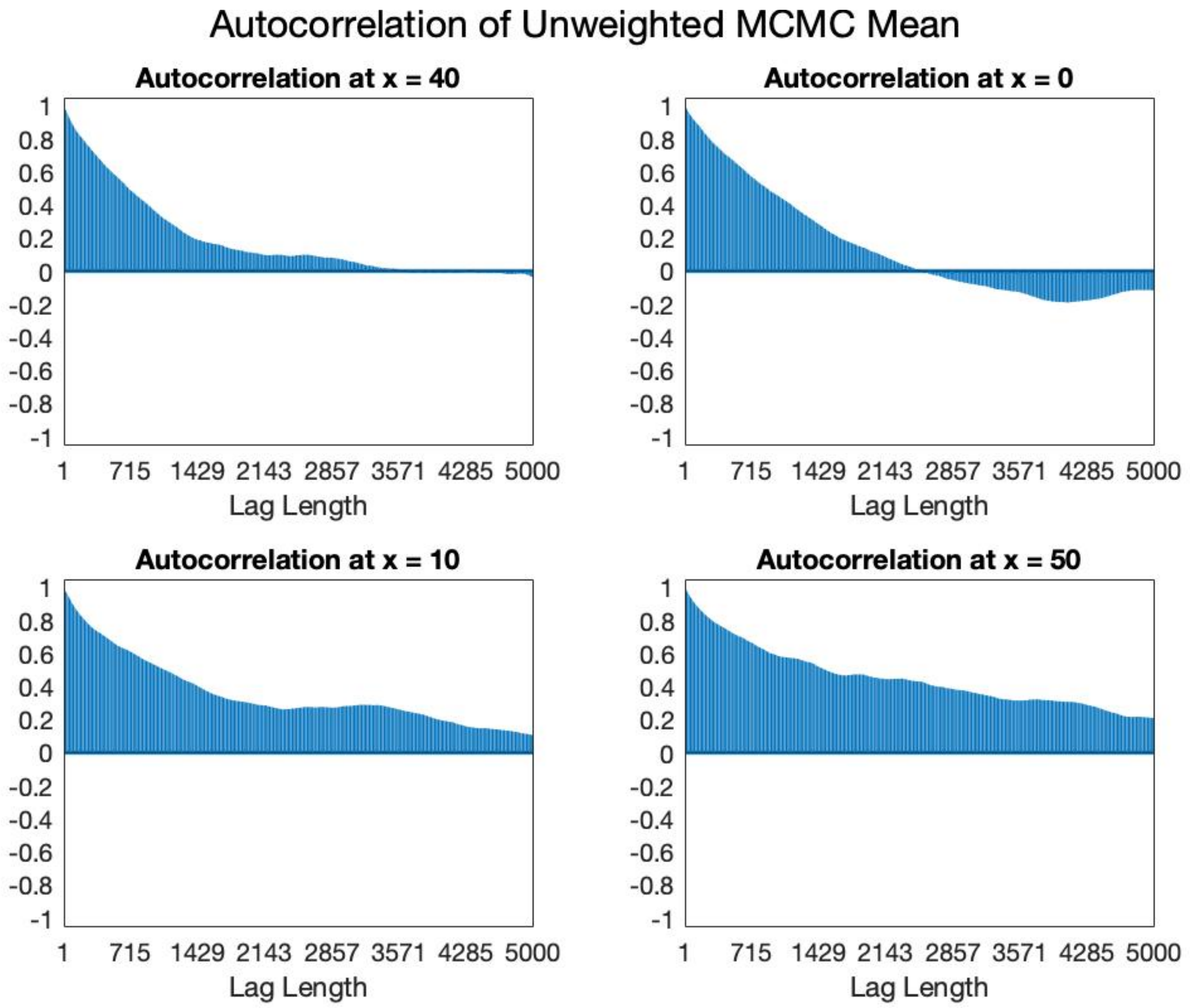


$\sigma = 0.75/\mathsf{SNR} = 27.43$

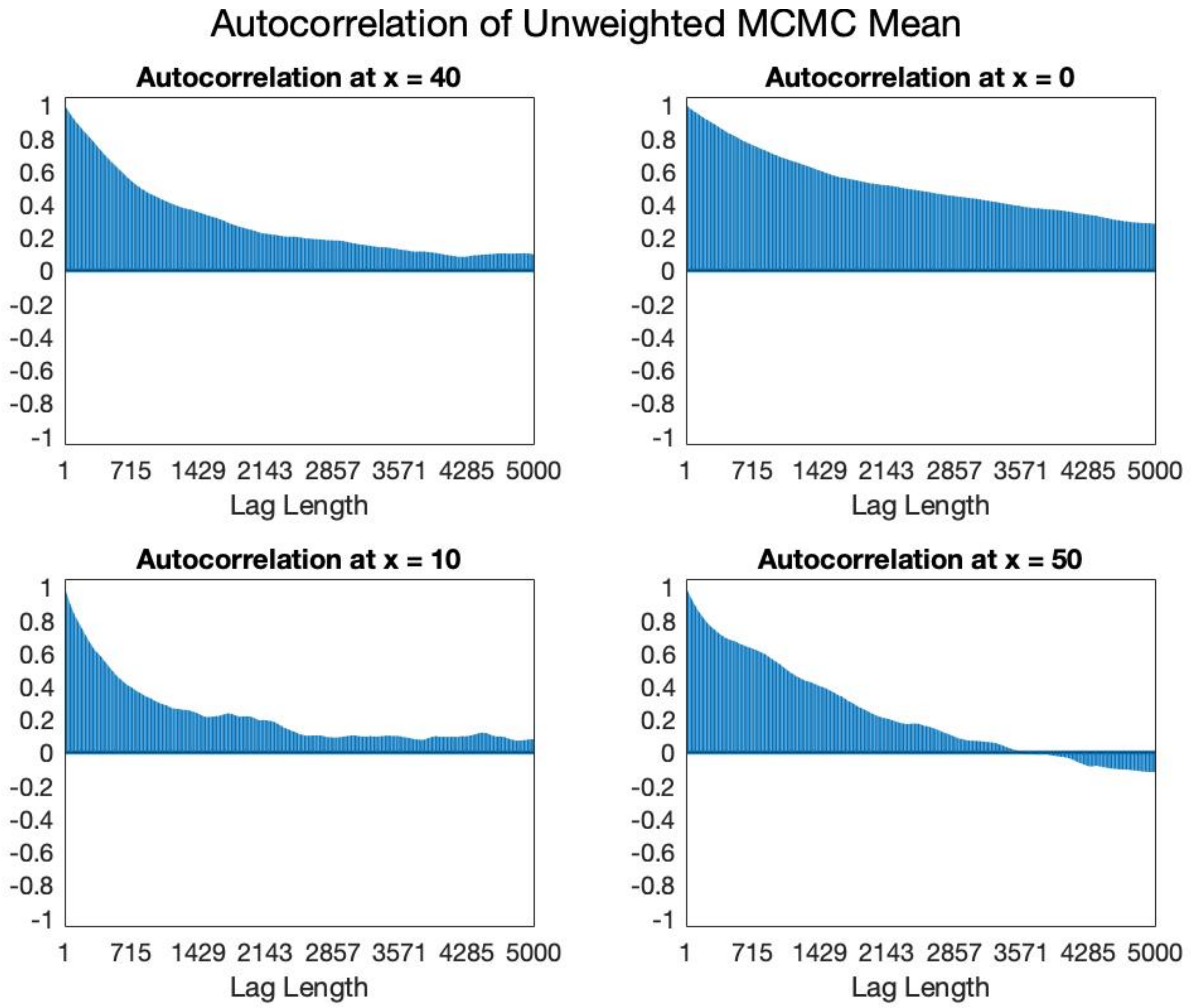


$\sigma = 1.00/\mathsf{SNR} = 25.13$

Unweighted MCMC

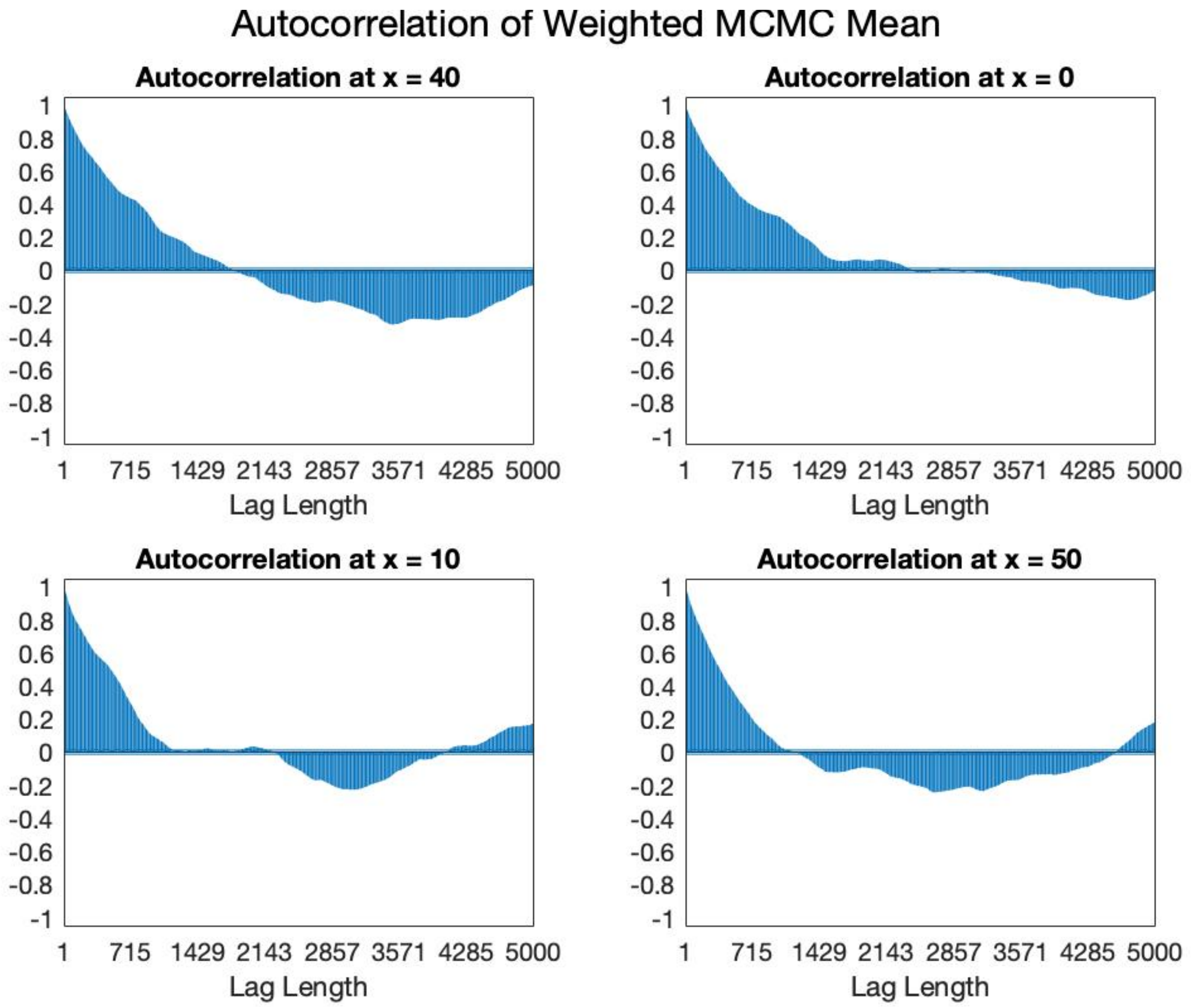


$\sigma = 0.25/\text{SNR} = 36.75$

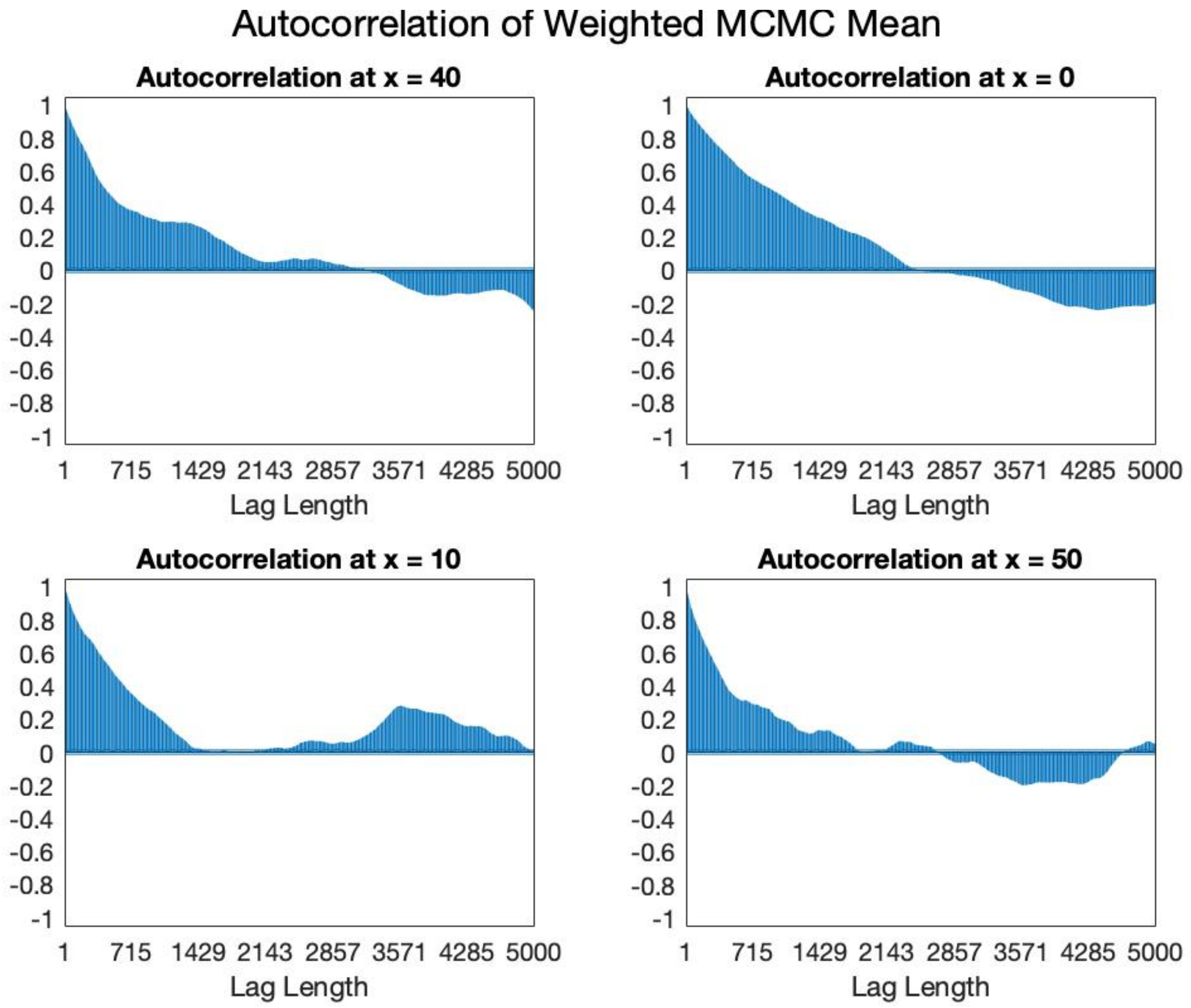


$\sigma = 1.00/\text{SNR} = 25.13$

Weighted MCMC

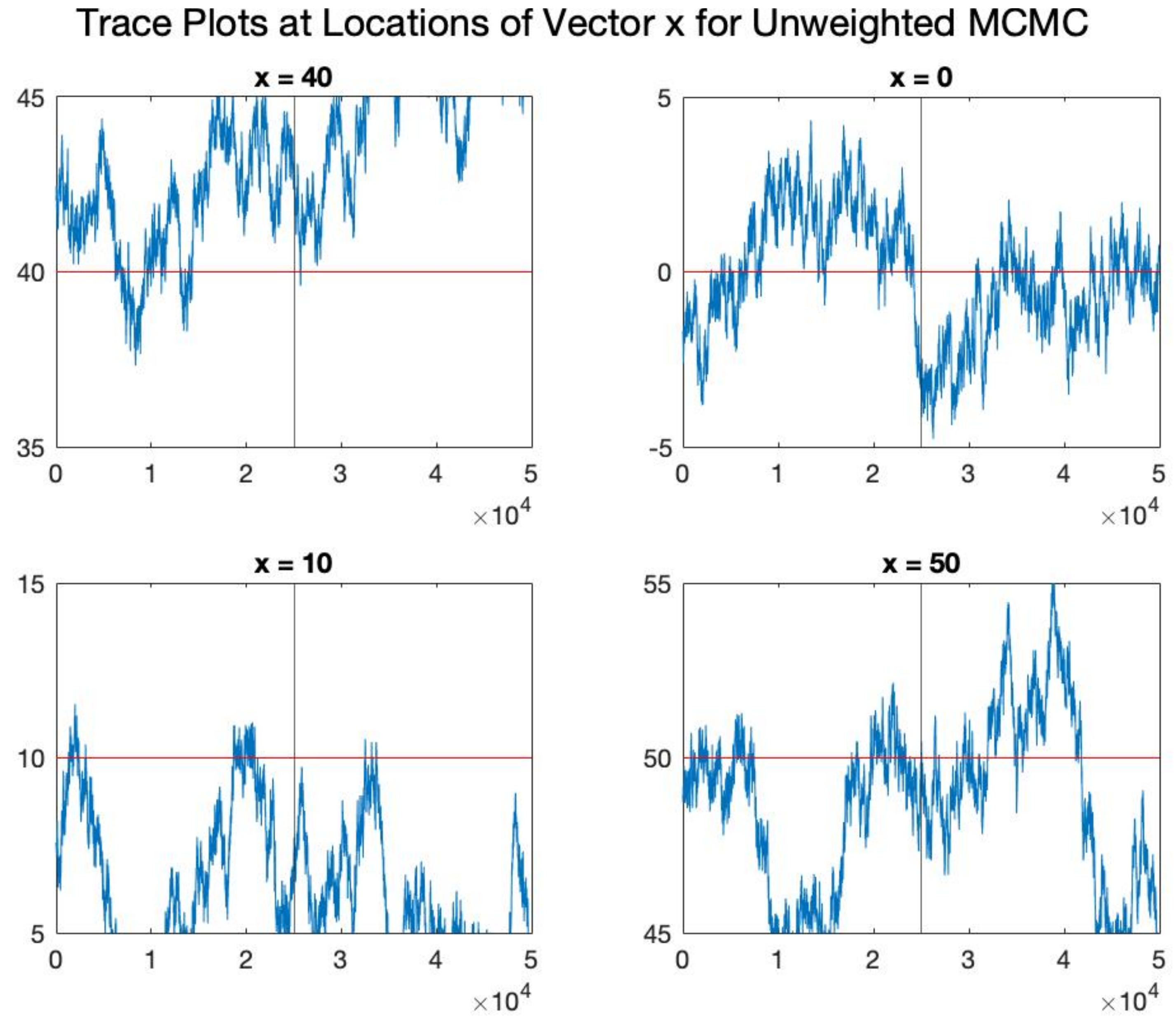


$\sigma = 0.25/\text{SNR} = 36.75$

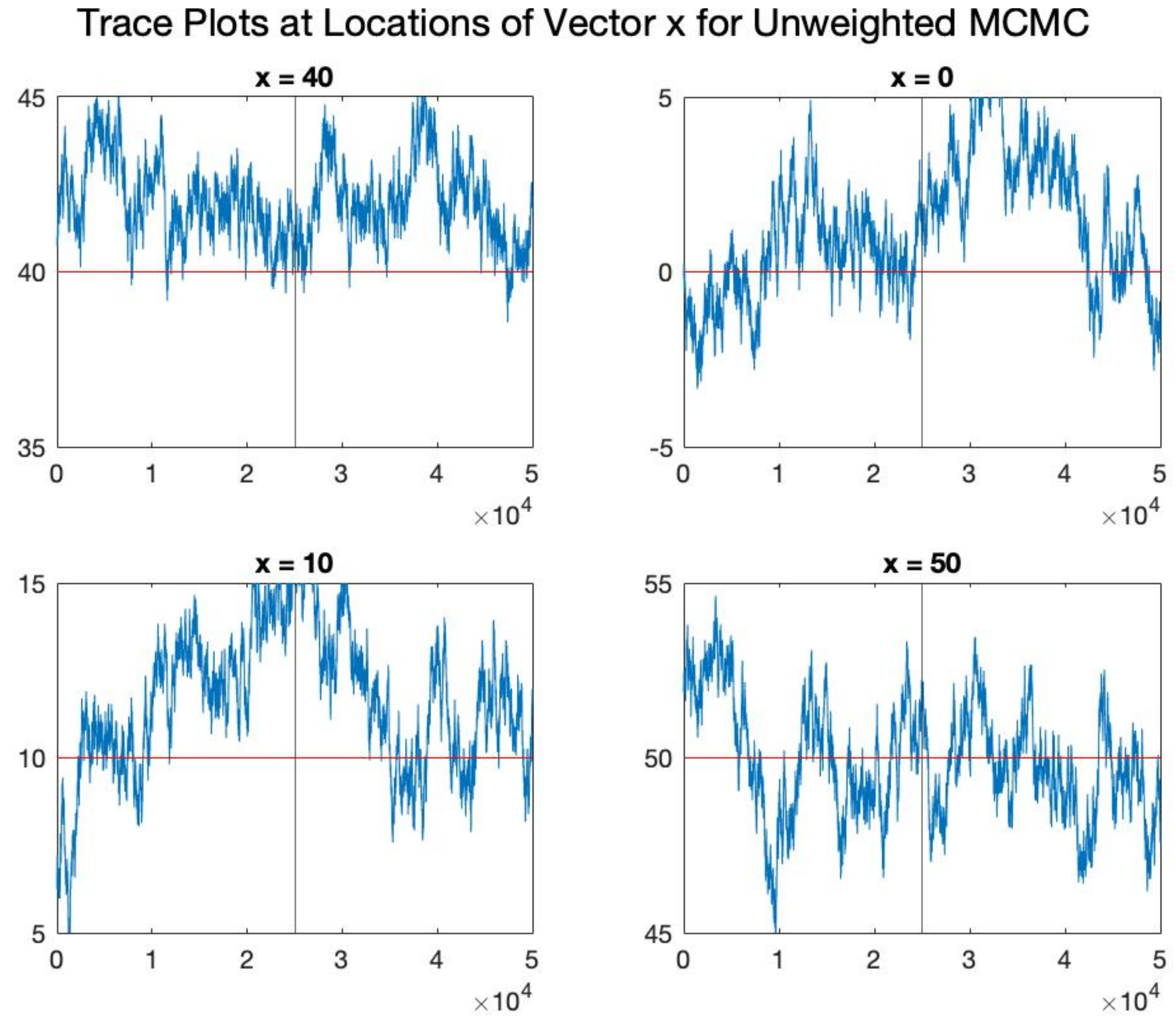


$\sigma = 1.00/\text{SNR} = 25.13$

Unweighted MCMC

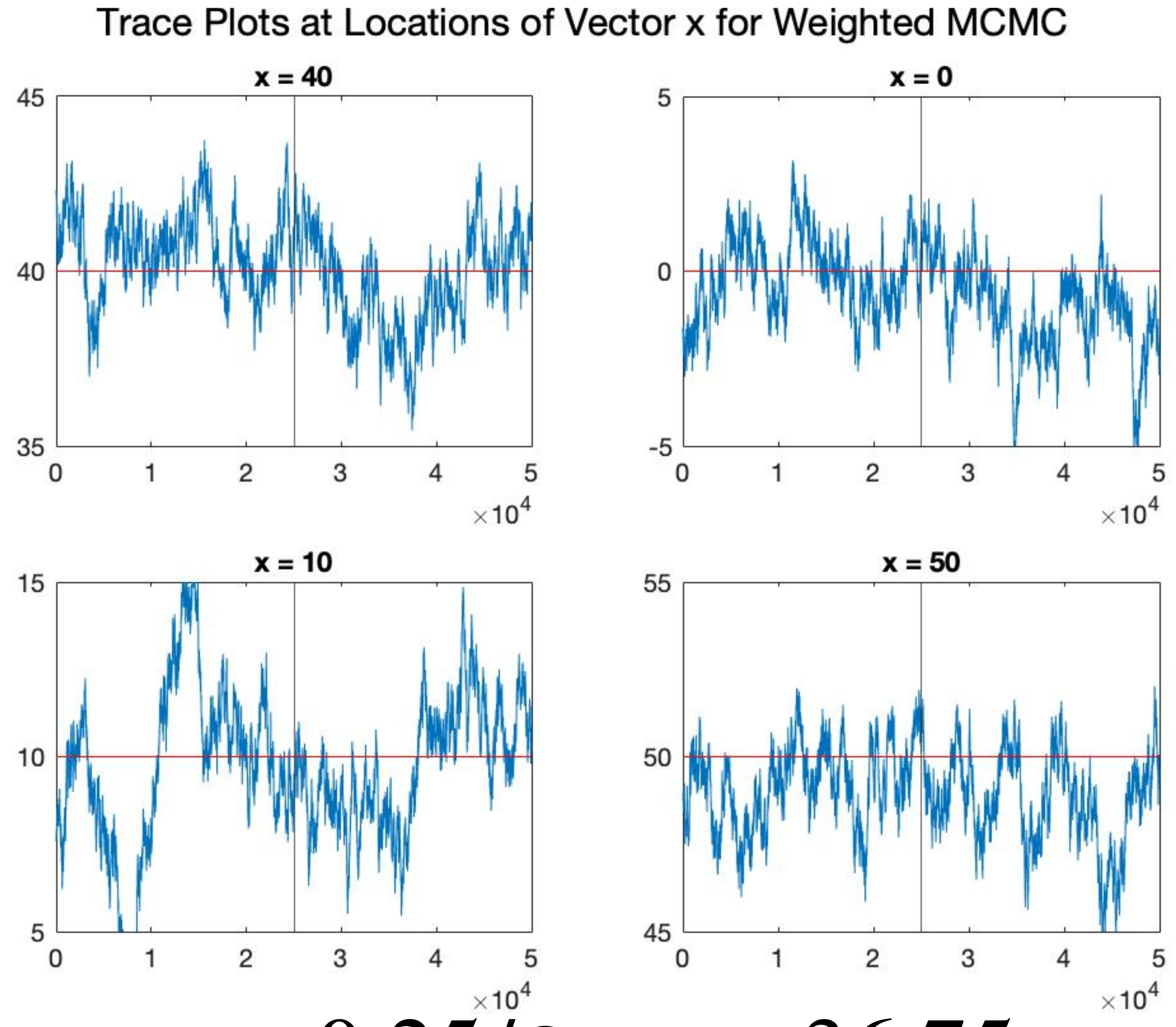


$\sigma = 0.25/\text{SNR} = 36.75$

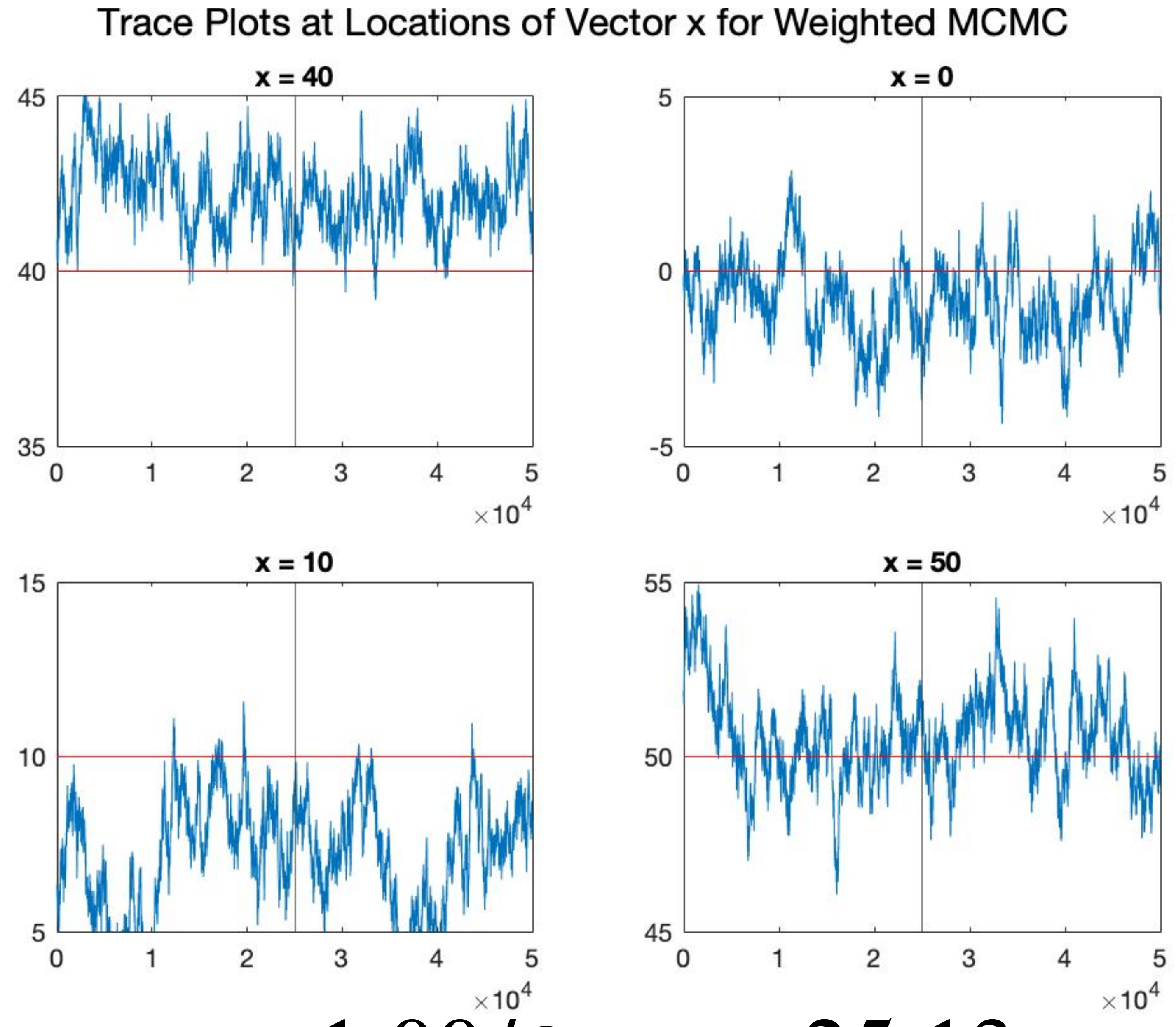


$\sigma = 1.00/\text{SNR} = 25.13$

Weighted MCMC



$\sigma = 0.25/\text{SNR} = 36.75$



$\sigma = 1.00/\text{SNR} = 25.13$

Conclusion

- The weighted MCMC gives better **mean recovery** than both the unweighted MCMC and just the MAP estimate.
- The regions where the **edge domain** is near-zero demonstrates the most improvements when weights are added to the prior.
- The **confidence intervals** are tighter for the weighted MCMC.
- The **auto-correlation** decays significantly faster with weights, which may indicate faster convergence to the posterior distribution.
- The **trace plots** at all regions show better convergence and more accurate solution recovery (but especially at regions without edge information).

Future Work

- The methodology in this thesis may be generalized to problems where the unknown is of **higher dimensions**, in which the convergence rate is even more important.

- The method to build weights with variance based joint sparsity may also consider first moment information such as the **expected value** of the regions in the sparse domain rather than just the variance.

- With the weights imposed in the prior, it may be now possible to use the ℓ_2 prior instead of the ℓ_1 to impose sparsity in the edge domain. The use of an ℓ_2 prior and a Gaussian likelihood function results in a Gaussian posterior by **conjugacy**. Then, we may use another MCMC technique named the **Gibbs sampler** to recover the unknown $\mathbf{X}$, which may improve both recovery accuracy and convergence rates.

Acknowledgement

I would like to express my sincere gratitude to Professor Anne Gelb for support me in not only in the work of this thesis but also my undergraduate experience in mathematics.

Further, I would also like to thank Dr. Theresa Scarnati at the Air Force Research Laboratory for all her mathematical expertise in this area of research and her guidance in the writing of this thesis.

Lastly, I would like to thank everyone for listening to my thesis presentation.

Selected References

- Ben Adcock, Anne Gelb, Guohui Song, and Yi Sui. Joint sparse recovery based on variances. SIAM Journal on Scientific Computing, 41(1):A246–A268, 2019.

- Rick Archibald, Anne Gelb, and Rodrigo Platte. Image reconstruction from undersampled fourier data using the polynomial annihilation transform. September 2011

- Johnathan M. Bardsley. Mcmc-based image reconstruction with uncertainty quantification. SIAM Journal on Scientific Computing, 34(3):A1316–A1332, 2012.

- Dani Gamerman and Hedibert F. Lopes. Morkov chain monte carlo. 68, 2006. MCMC.

- Anne Gelb and Theresa Scarnati. Reducing effects of bad data using variance based joint sparsity recovery. Journal of Scientific Computing, 78(1):94–120, 2019.

Selected References

- W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. Biometrika, 57(1):97–109, 04 1970.

- J. Kaipio and E. Somersalo. Statistical and Computational Inverse Problems. Applied Mathematical Sciences. Springer New York, 2006.

- Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller. Equation of state calculations by fast computing machines. The Journal of Chemical Physics, 21(6):1087–1092, 1953.

- C. Robert and G. Casella. Introducing Monte Carlo Methods with R. Use R! Springer, 2010.

- Theresa Scarnati. Inverse problems in a bayesian framework, December 2019.