Their Setting

Suppose we have n samples of multi-response data $\left\{(\mathbf{X}_i, \mathbf{y}_i)\right\}_{i=1}^{n}$

where each observation is of the form $\mathbf{y}_i = \mathbf{X}_i\theta_* + \eta_i, \;\; \eta_i = \Sigma_*^{1/2}\tilde{\eta}_i$

where $\mathbf{y}_i \in \mathbb{R}^m, \quad \mathbf{X}_i \in \mathbb{R}^{m\times p}$

and $\tilde{\eta}_i \in \mathbb{R}^m$ is zero-mean isotropic noise

Method: Estimate θ_* and Σ_* from the data using alternating minimization (AltMin):

$$\hat{\Sigma}_{(t+1)} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - X_i\hat{\theta}_{(t)}\right)\left(y_i - X_i\hat{\theta}_{(t)}\right)^T$$

$$\hat{\theta}_{(t+1)} = arg\min_{\theta\in\mathbb{R}^p}\frac{1}{2n}\sum_{i=1}^{n}\|\hat{\Sigma}_{(t+1)}^{-\frac{1}{2}}(y_i - X_i\theta)\|_2^2 \quad \textbf{s.t.} \quad f(\theta)\le\lambda$$

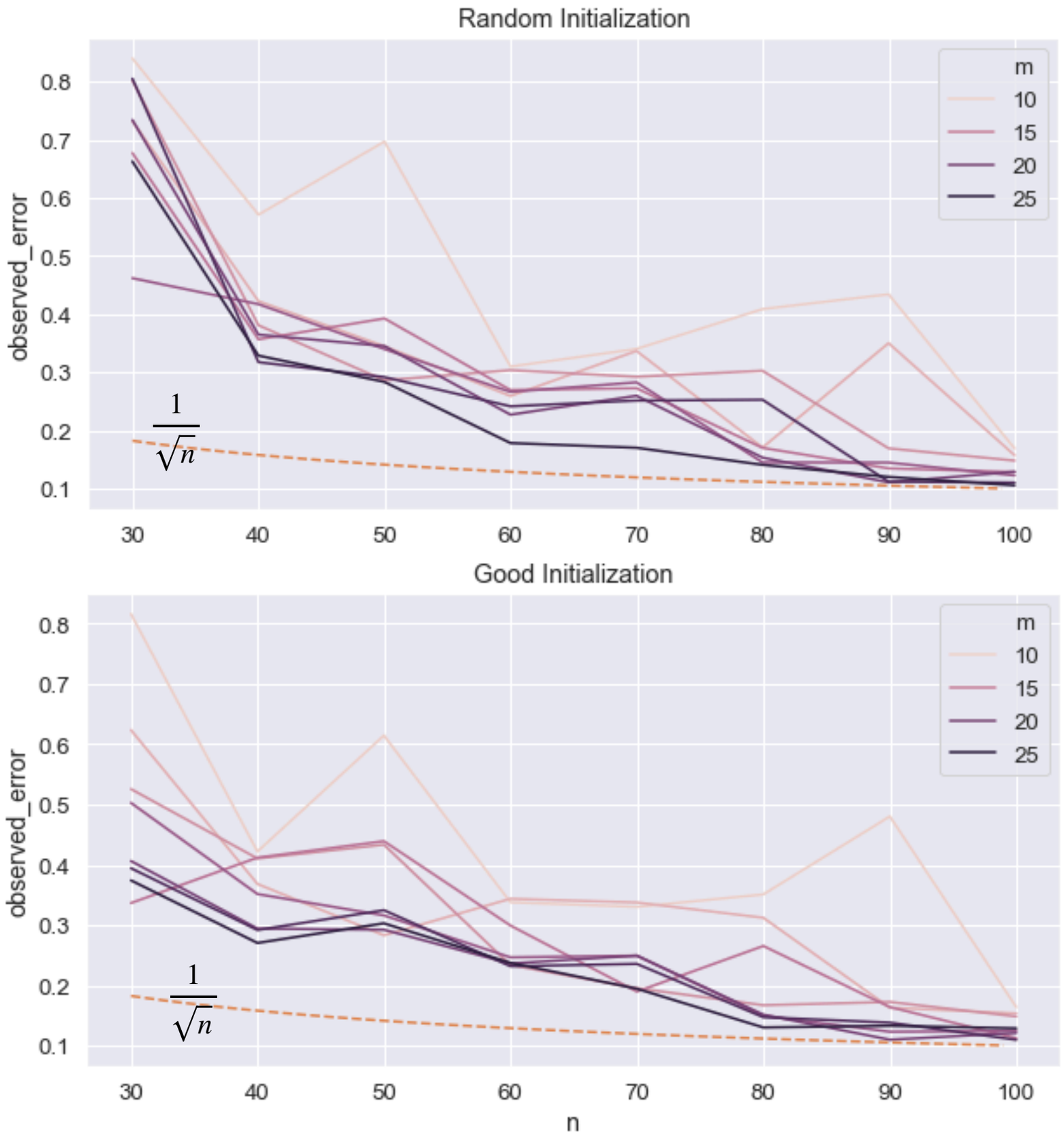
Error: $\|\hat{\theta}_{(T)} - \theta_*\| \le e_{\min} + \rho_n^T\left(\|\hat{\theta}_{(0)} - \theta_*\|_2 - e_{\min}\right)$

$e_{\min} = O\left(\frac{w(C)+m}{\sqrt{n}}\right)$ arbitrary initialization

$e_{\min} = O\left(\frac{w(C)}{\sqrt{n}}\right)$ good initialization

$\rho_n < 1$ contraction factor

Q: do we observe the theoretical errors in practice?



$$f(\theta) = \|\theta\|_2^2$$

$$p = 100$$

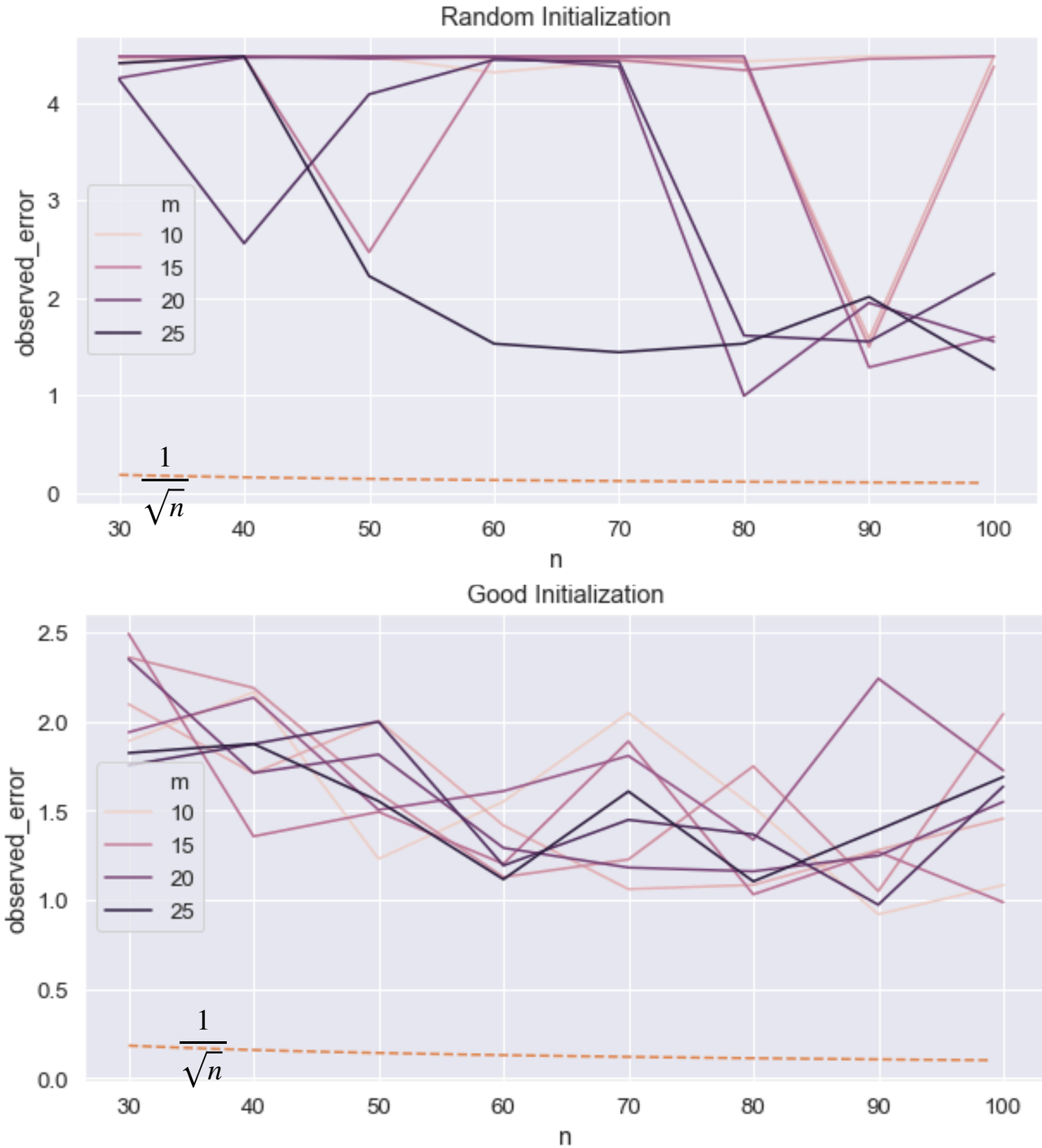
$$\Sigma_* = diag\left(\begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}, \ldots \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}\right) \in \mathbb{R}^{m \times m}$$

$$X_1, \ldots X_n \sim N(0, I) \in \mathbb{R}^{m \times p}$$

$$\theta_* = \begin{bmatrix} 1 \\ \vdots \\ 1 \\ -1 \\ \vdots \\ -1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \begin{matrix} \} \ 10 \\ \} \ 10 \\ \} \ p - 20 \end{matrix}$$

$$\mathbf{y}_i = \mathbf{X}_i \theta_* + \eta_i, \quad \eta_i = \Sigma_*^{1/2} \tilde{\eta}_i$$

Q: do we observe the theoretical errors in practice?



$$f(\theta) = \|\theta\|_1$$

$$p = 100$$

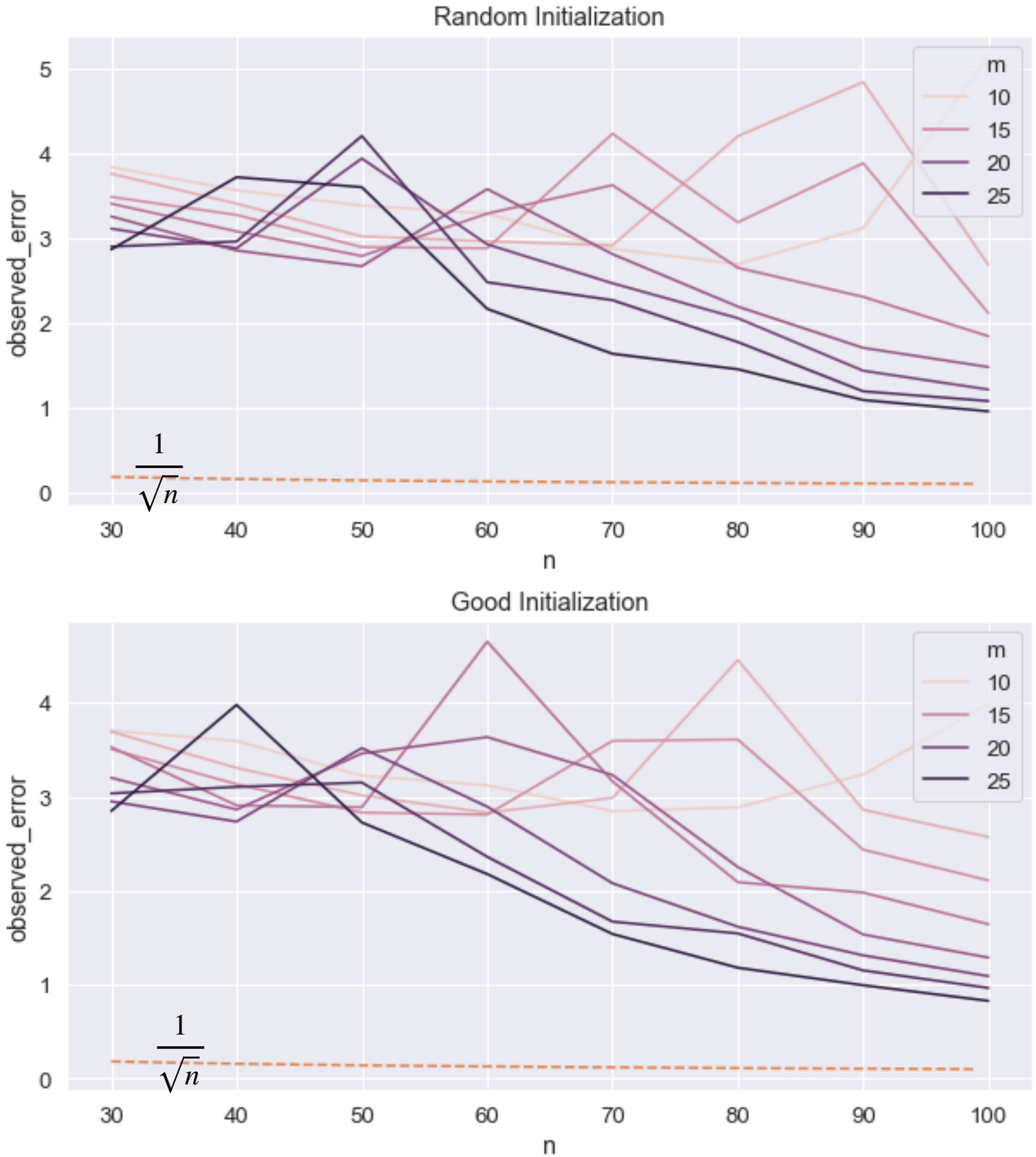
$$\Sigma_* = diag\left(\begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}, \ldots \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}\right) \in \mathbb{R}^{m \times m}$$

$$X_1, \ldots X_n \sim N(0, I) \in \mathbb{R}^{m \times p}$$

$$\theta_* = \begin{bmatrix} 1 \\ \vdots \\ 1 \\ -1 \\ \vdots \\ -1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \begin{matrix} \} \ 10 \\ \} \ 10 \\ \} \ p - 20 \end{matrix}$$

$$\mathbf{y}_i = \mathbf{X}_i \theta_* + \eta_i, \quad \eta_i = \Sigma_*^{1/2} \tilde{\eta}_i$$

Q: do we observe the theoretical errors in practice?



$$f(\theta) = \|\theta\|_2^2$$

$$p = 1000$$

$$\Sigma_* = diag\left(\begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}, \ldots \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}\right) \in \mathbb{R}^{m \times m}$$

$$X_1, \ldots X_n \sim N(0, I) \in \mathbb{R}^{m \times p}$$

$$\theta_* = \begin{bmatrix} 1 \\ \vdots \\ 1 \\ -1 \\ \vdots \\ -1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \begin{matrix} \} \ 10 \\ \} \ 10 \\ \} \ p - 20 \end{matrix}$$

$$\mathbf{y}_i = \mathbf{X}_i \theta_* + \eta_i, \quad \eta_i = \Sigma_*^{1/2} \tilde{\eta}_i$$

Our Setting

We have observations and simulations, as follows:

Observations

$$X_{obs} \in \mathbb{R}^{n_o \times p}, \quad y_{obs} \in \mathbb{R}^{n_o}$$

$$y_{obs} = X_{obs}\beta + \epsilon_o$$

$$\epsilon_o \sim N(0,\sigma^2 I)$$

Simulations

$$X_{sim} \in \mathbb{R}^{n_s \times p}, \quad y_{sim} \in \mathbb{R}^{n_s}$$

$$\boxed{n_s = kn_o}$$

$$\begin{aligned} y_{sim} &= X_{sim}(\beta + \Delta) + \epsilon_s \\ &= X_{sim}\beta + (X_{sim}\Delta + \epsilon_s) \\ &= X_{sim}\beta + \eta \end{aligned}$$

$$\epsilon_s \sim N(0,\sigma^2 I)$$

$$\Delta \sim N(0,\Sigma_\Delta)$$

$$\eta \sim N(0,\tilde{\Sigma})$$

$$\tilde{\Sigma} = X_{sim}\Sigma_\Delta X_{sim}^T + \sigma^2 I$$

Questions:

How does our setting relate to theirs?

What do their error bounds look like in our setting?

How can we leverage the specific structure of our setting to improve the error bounds?

Option 1: Stack observations and simulations

$$\begin{bmatrix} y_{obs} \\ y_{sim} \end{bmatrix} = \begin{bmatrix} X_{obs} \\ X_{sim} \end{bmatrix} \theta_* + \Sigma_*^{1/2} \tilde{\eta} \qquad \Sigma_* = \begin{bmatrix} \sigma^2 I_{n_o} & 0 \\ 0 & \tilde{\Sigma} \end{bmatrix} \qquad \tilde{\Sigma} = X_{sim} \Sigma_\Delta X_{sim}^T + \sigma^2 I_{n_s}$$

$$n_s = k n_o$$

Error

$$e_{\min} = O\left(\frac{w(C) + (k+1)n_o}{\sqrt{1}} \right)$$

arbitrary initialization

$$e_{\min} = O\left(\frac{w(C)}{\sqrt{1}} \right)$$

good initialization

Option 2: Make groups of observations and simulations of size k

We are assuming $n_s = kn_o$ (i.e. the number of simulated data points is some constant k times the number of observed data points), we can alternatively group our data as follows:

$$y_{obs} = \begin{bmatrix} y_{obs,1} \\ \vdots \\ y_{obs,n_o} \end{bmatrix} \qquad X_{obs} = \begin{bmatrix} X_{obs,1} \\ \vdots \\ X_{obs,n_o} \end{bmatrix}$$

$$y_{sim} = \begin{bmatrix} y_{sim,1,1} \\ \vdots \\ y_{sim,1,k} \\ \vdots \\ y_{sim,n_o,k} \end{bmatrix} \qquad X_{sim} = \begin{bmatrix} X_{sim,1,1} \\ \vdots \\ X_{sim,1,k} \\ \vdots \\ X_{sim,n_o,k} \end{bmatrix}$$

Then we let

$$\mathbf{y}_i = \begin{bmatrix} y_{obs,i} \\ y_{sim,i,1} \\ \vdots \\ y_{sim,i,k} \end{bmatrix} \in \mathbb{R}^{k+1} \quad \& \quad \mathbf{X}_i = \begin{bmatrix} X_{obs,i} \\ X_{sim,i,1} \\ \vdots \\ X_{sim,i,k} \end{bmatrix} \in \mathbb{R}^{k+1\times p}$$

and so our data are $\left\{ (\mathbf{X}_i, \mathbf{y}_i) \right\}_{i=1}^{n_o}$

where $\mathbf{y}_i = \mathbf{X}_i\theta_* + \Sigma_*^{1/2}\tilde{\eta}$

$$\Sigma_* = \begin{bmatrix} \sigma^2 & 0 \\ 0 & \mathbf{X}_i\Sigma_\Delta\mathbf{X}_i^T + \sigma^2 I_k \end{bmatrix}$$

Error

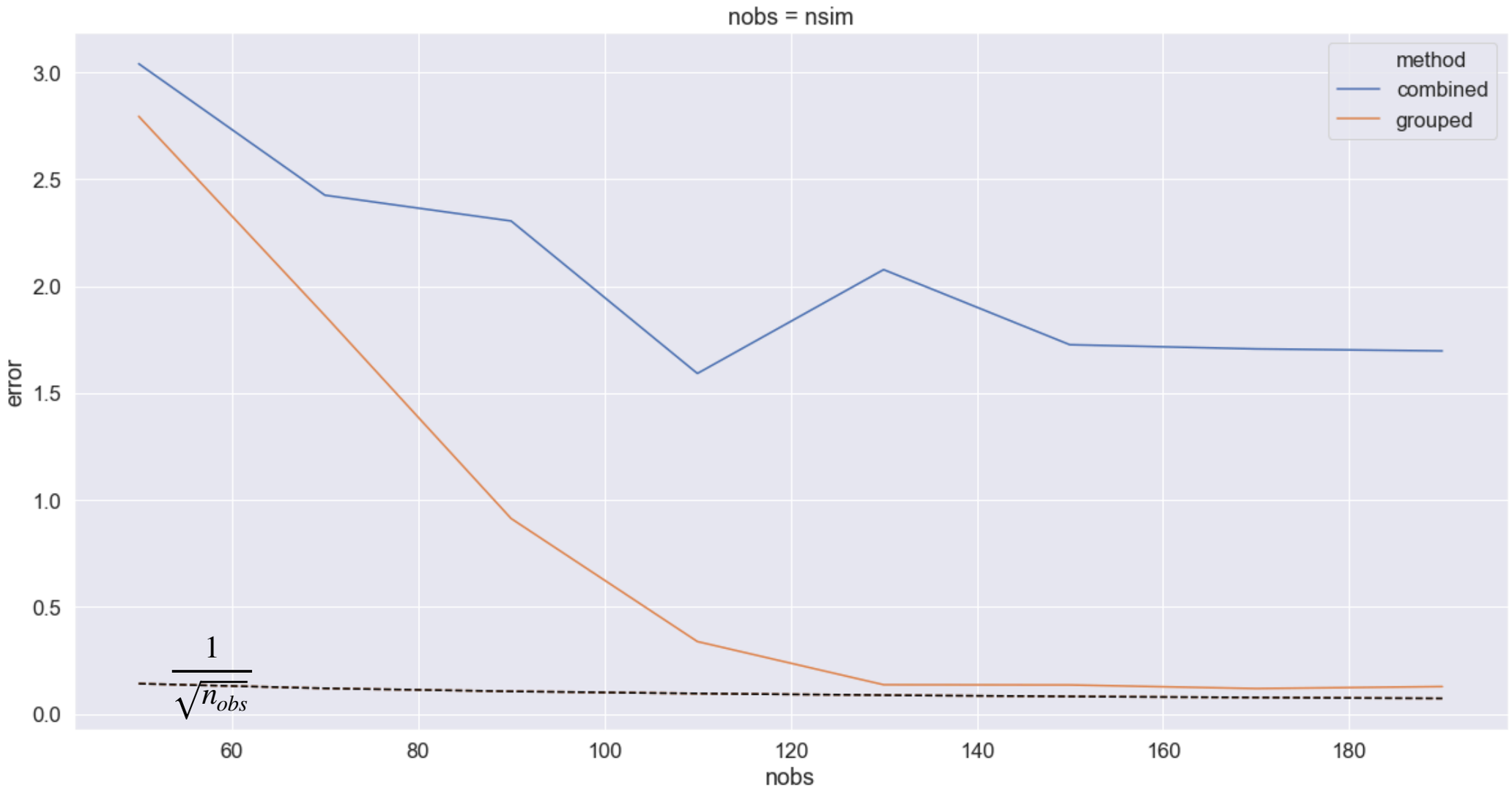
$$e_{\min} = O\left(\frac{w(C) + (k+1)}{\sqrt{n_o}}\right)$$

arbitrary initialization

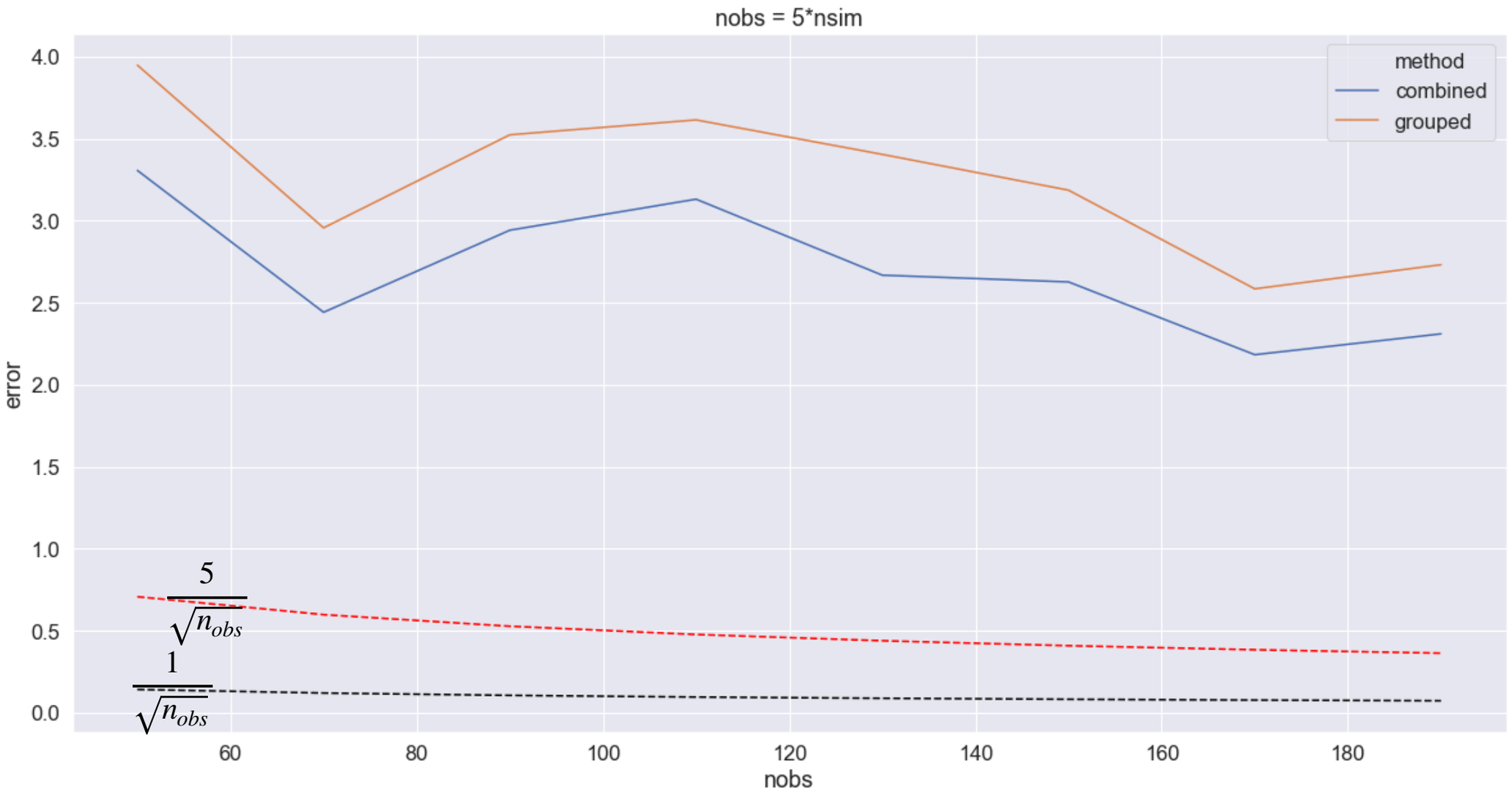
$$e_{\min} = O\left(\frac{w(C)}{\sqrt{n_o}}\right)$$

good initialization

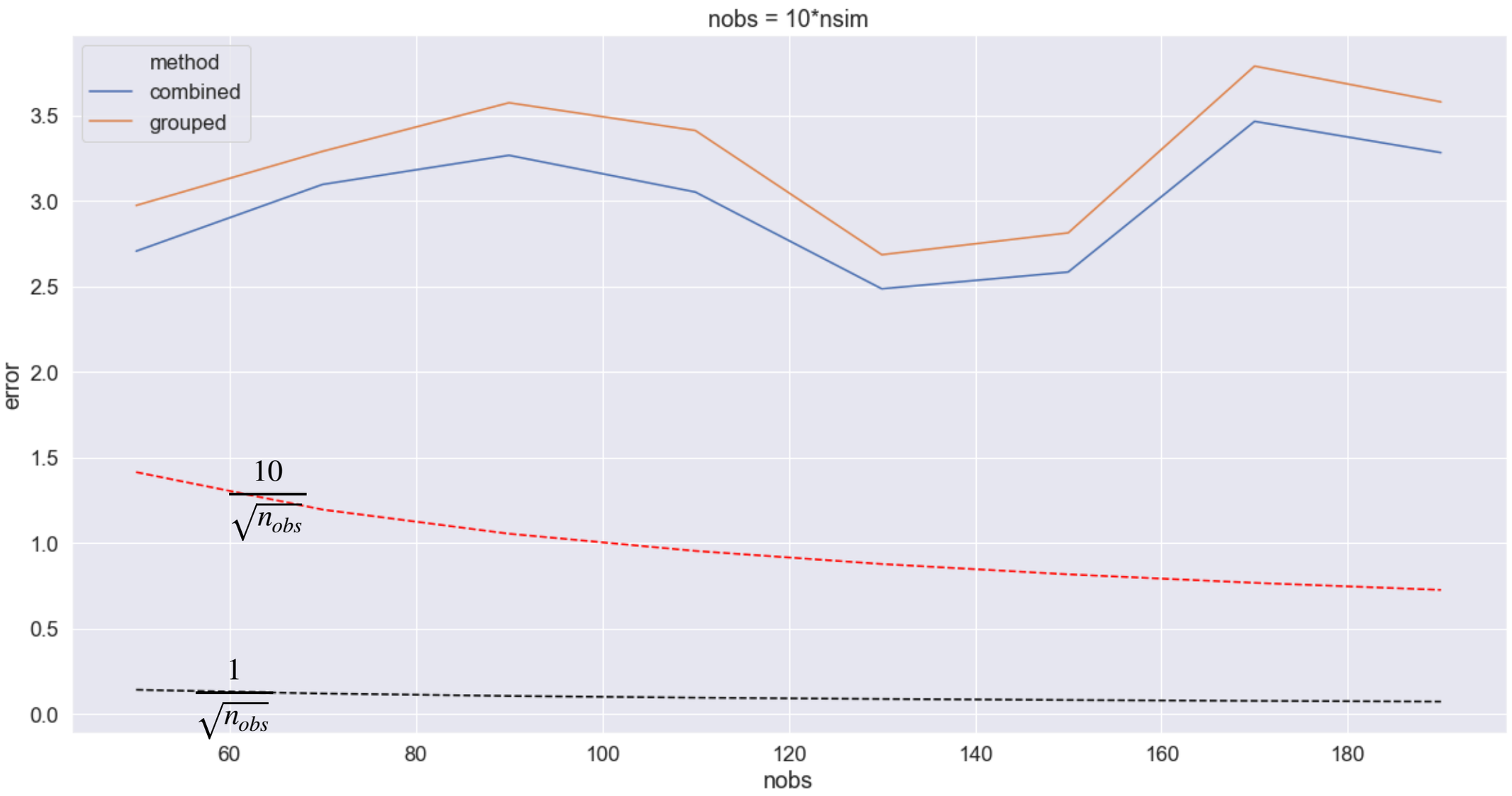
Q1: how do these different groupings affect the errors?



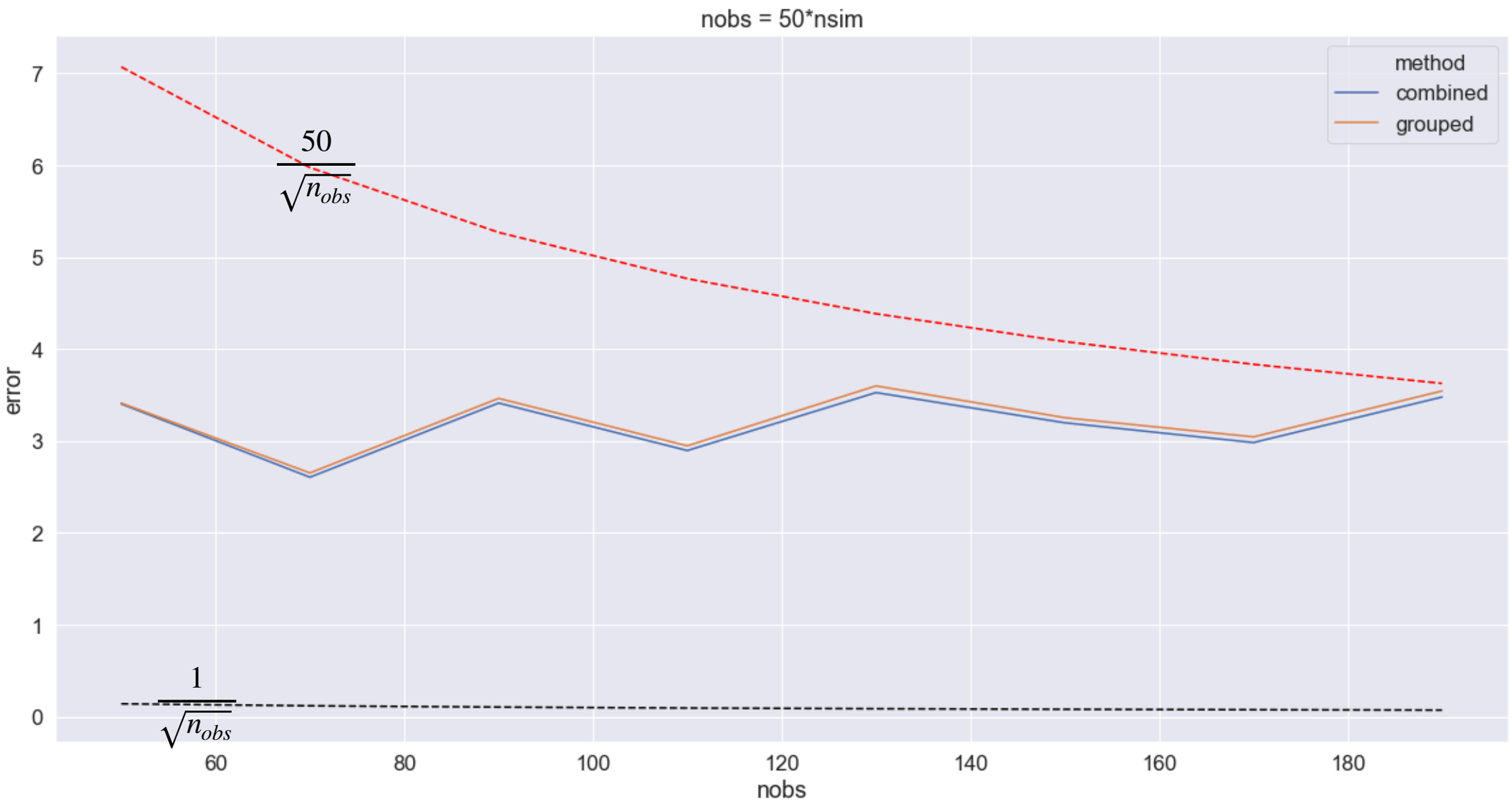
Q1: how do these different groupings affect the errors?



Q1: how do these different groupings affect the errors?



Q1: how do these different groupings affect the errors?

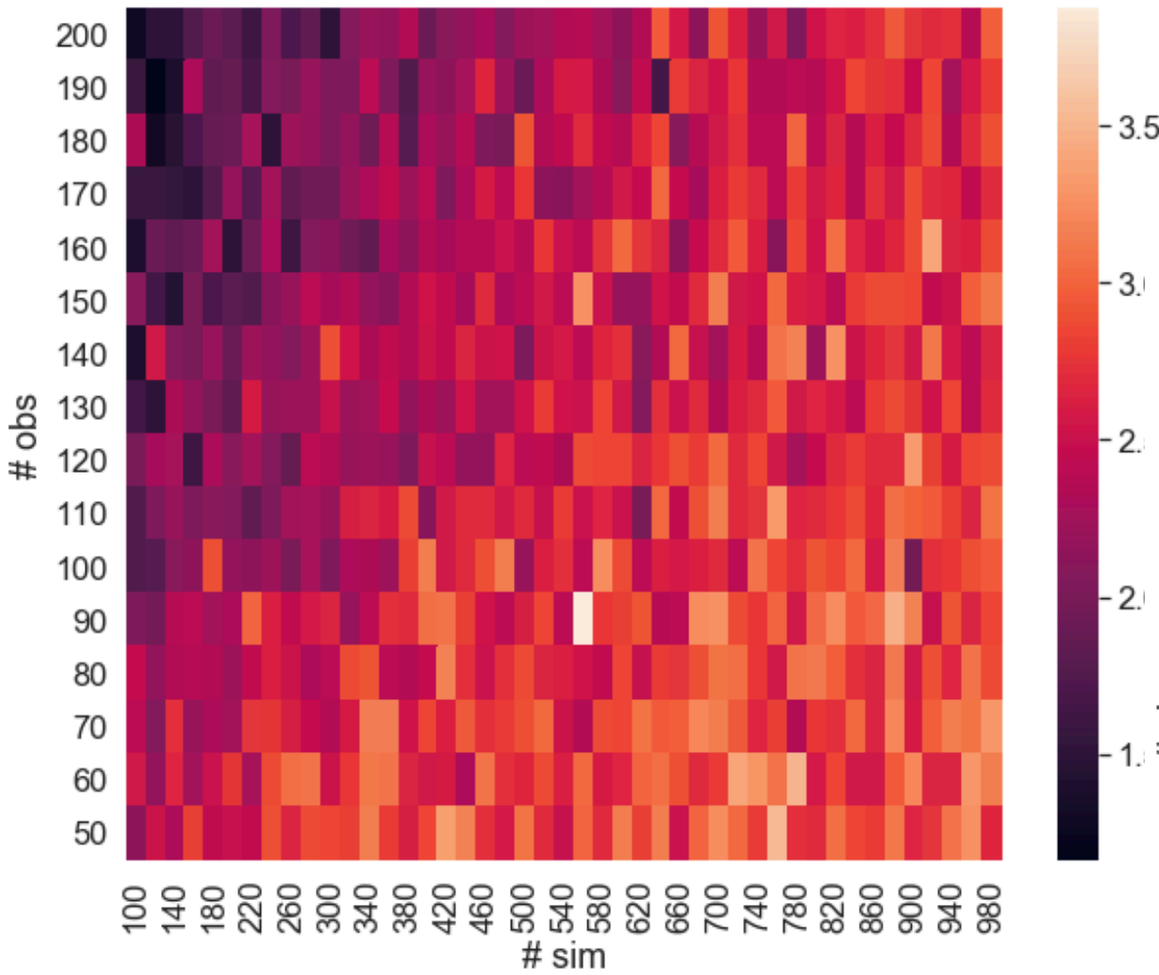


Q2: can we exploit covariance structure to get a better estimate?

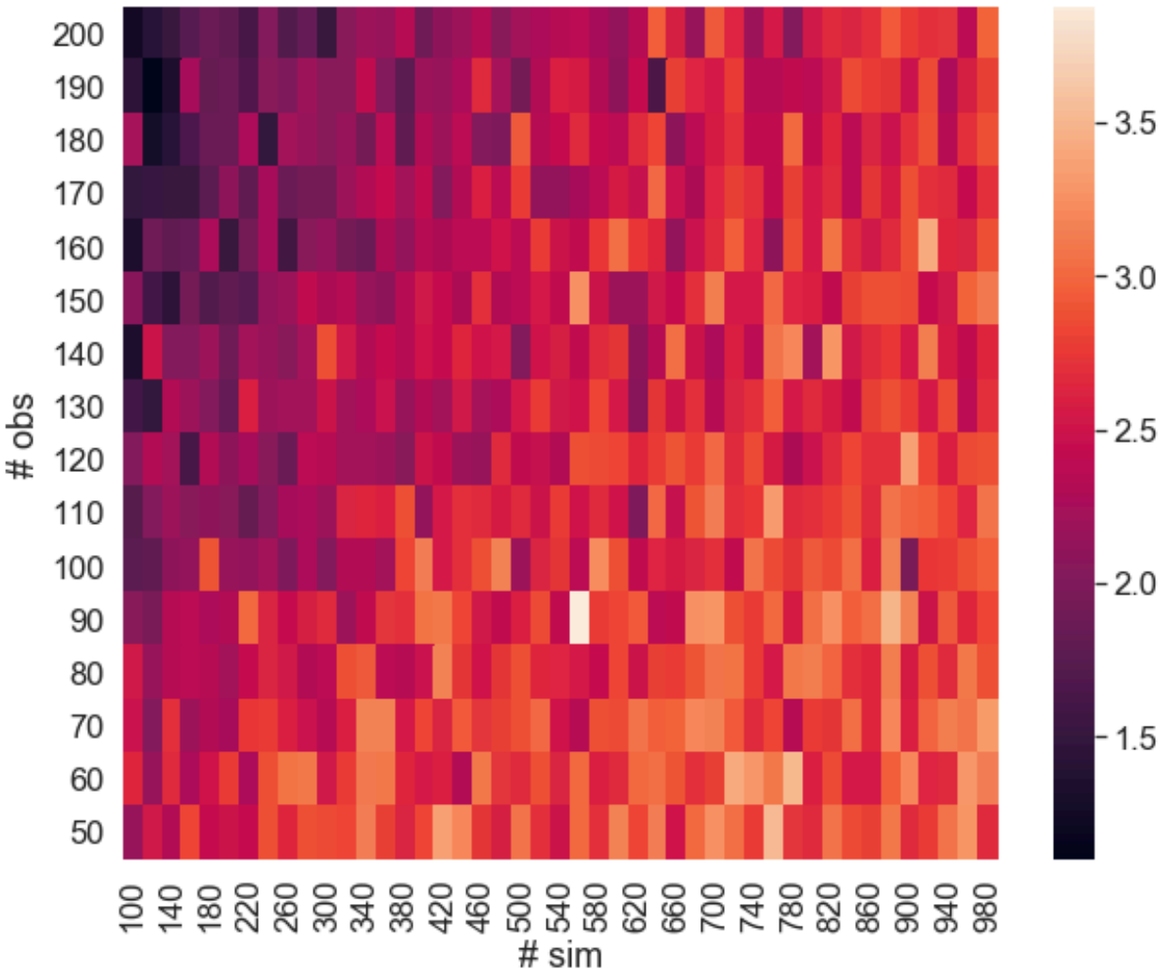
They assume no structure on Σ_* while we assume $\Sigma_* = \begin{bmatrix} \sigma^2 I_{n_o} & 0 \\ 0 & \tilde{\Sigma} \end{bmatrix}$

$$\tilde{\Sigma} = X_{sim}\Sigma_\Delta X_{sim}^T + \sigma^2 I_{n_s}$$

no structure assumed



structure exploited



Surprisingly, these are nearly identical. I will try a few different experimental settings before drawing any conclusions here.

Q: can we use observations to initialize the procedure and simulations to train?

i.e. $\theta_0 = argmin_\theta \left\{ \|y_{obs} - X_{obs}\theta\| + \lambda \|\theta\|_2^2 \right\}$

$\hat{\theta}$ is the result of running AltMin on (X_{sim}, y_{sim}) initialized at θ_0

simulation setting: $X_{obs} \sim N(0, I_{n_o})$

$y_{obs} = X_{obs}\theta^* + \epsilon_{obs} \qquad \epsilon_{obs} \sim N(0, 0.1)$

$$\theta^* = \begin{bmatrix} 1 \\ \vdots \\ 1 \\ -1 \\ \vdots \\ -1 \\ 0 \\ \vdots \\ \vdots \\ 0 \end{bmatrix}$$

$X_{sim} \sim N(0, I_{n_s})$

$y_{sim} = X_{sim}(\theta^* + \Delta) + \epsilon_{sim}$

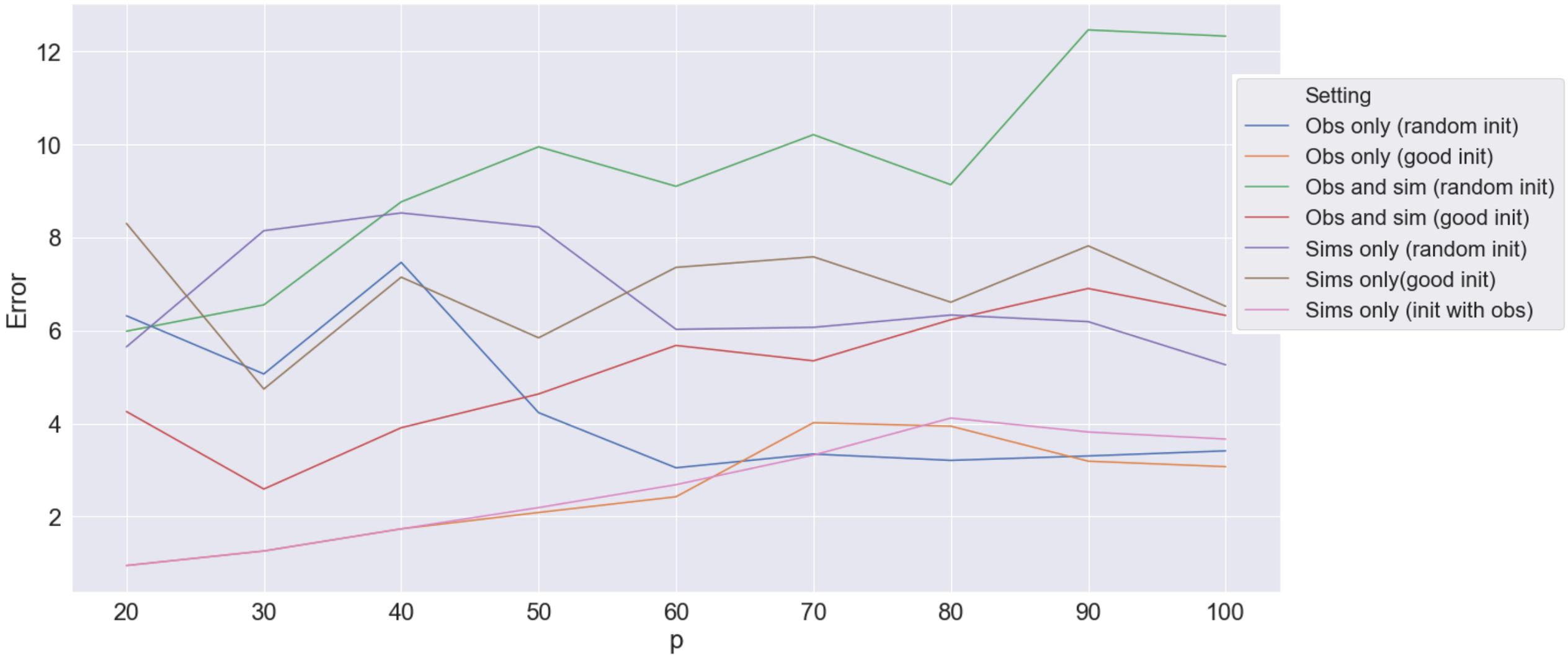
$\epsilon_{sim} \sim N(0, 0.1)$

$\Delta \sim N(0, \Sigma_\Delta)$

$\Sigma_\Delta = diag\left(\begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}, \dots \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix} \right) \in \mathbb{R}^{n_s \times n_s}$

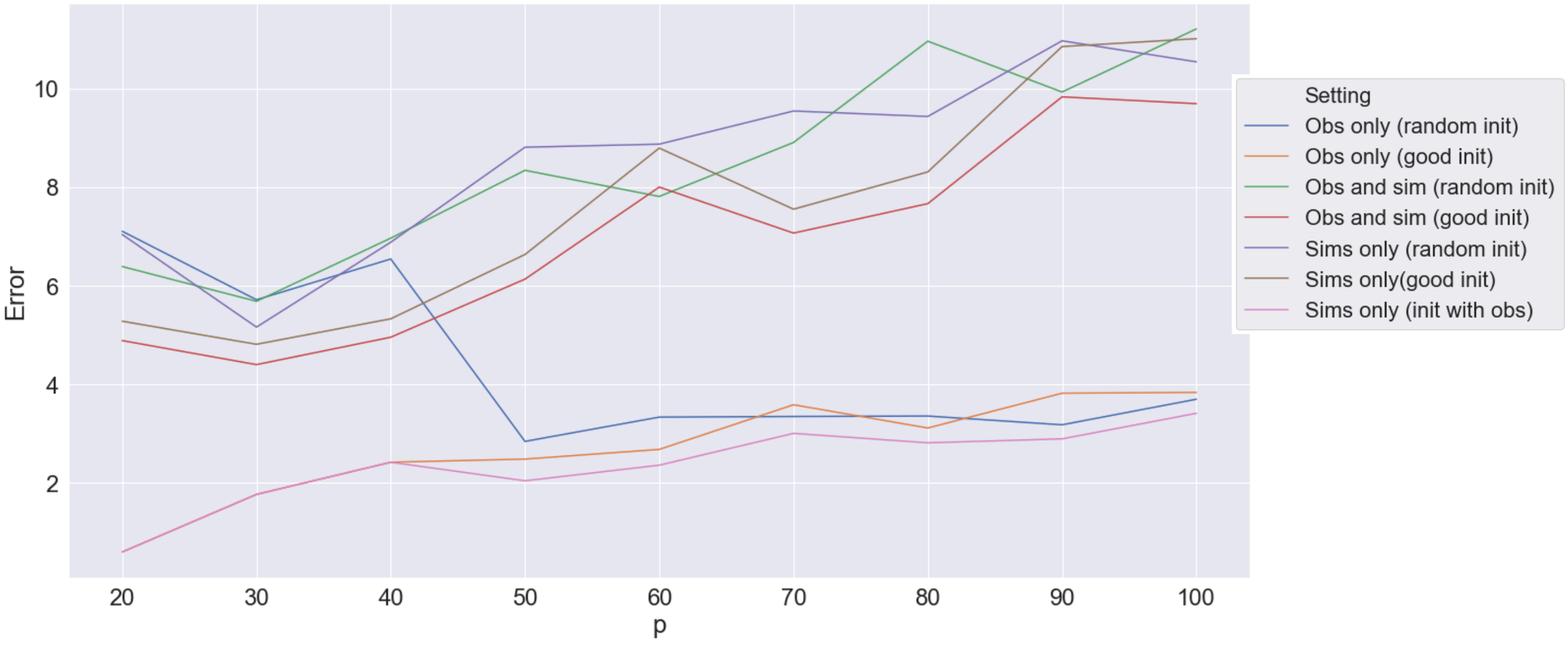
Simulation: initialize with observations

obs = # sim = 50



Simulation: initialize with observations

obs = 50, # sim = 500



Simulation: initialize with observations

obs = 50, # sim = 2500