COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

Bias-Free Scalable Gaussian Processes via Randomized Truncations

Andres Potapczynski *, Luhuan Wu *, **Dan Biderman** *, Geoff Pleiss, John Cunningham

Gaussian Process hyperparameter learning

argmin $\theta \longrightarrow$ **Loss = model complexity + fitting error**

model complexity: $\log \det(\mathbf{K}_\theta)$

fitting error: $\mathbf{K}_\theta^{-1}\mathbf{y}$

$K_\theta : N \times N$

scalable approximations

savings:
using $J \ll N$
Fourier features

Random Fourier Features (RFF)

$$\mathbf{K} \approx \sum_{j=1}^{J} \boldsymbol{\phi}_j \boldsymbol{\phi}_j^\top$$

Rahimi and Recht (2008)

Conjugate Gradients (CG)

$$\mathbf{K}^{-1}\mathbf{y} \approx \sum_{j=1}^{J} \gamma_j \mathbf{d}_j$$

Cunningham et al., 2008, Cutajar et al., 2016,
Gardner, Pleiss et al., 2018

savings:
early stopping at
iteration $J \ll N$

How do early-truncation procedures affect GP learning?

Loss = model complexity + fitting error

computation VS bias

scalable approximations

savings:
using $J \ll N$
Fourier features

Random Fourier Features (RFF)

$$\mathbf{K} \approx \sum_{j=1}^{J} \phi_j \phi_j^\top$$

Rahimi and Recht (2008)

Conjugate Gradients (CG)

$$\mathbf{K}^{-1}\mathbf{y} \approx \sum_{j=1}^{J} \gamma_j \mathbf{d}_j$$

Cunningham et al., 2008, Cutajar et al., 2016,
Gardner, Pleiss et al., 2018

savings:
early stopping at
iteration $J \ll N$

Thm 1: CG underfits the data

Loss = model complexity + fitting error

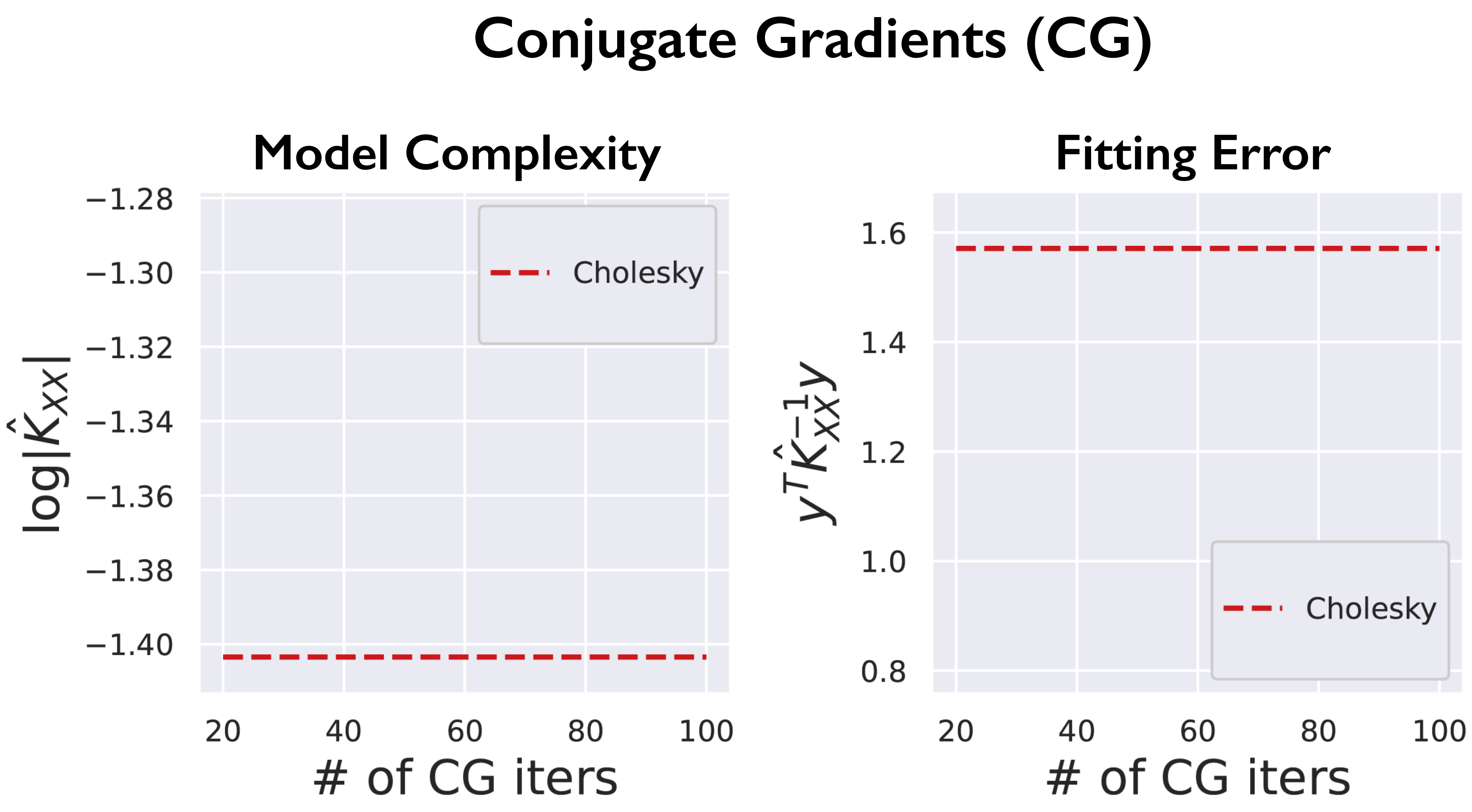
Theorem 1. *Let u_J and v_J be the estimates of $\mathbf{y}^\top \widehat{\mathbf{K}}_{\mathbf{XX}}^{-1}\mathbf{y}$ and $\log|\widehat{\mathbf{K}}_{\mathbf{XX}}|$ respectively after J iterations of CG; i.e.:*

$$u_J = \mathbf{y}^\top \left(\sum_{i=1}^{J} \gamma_i \mathbf{d}_i\right), \quad v_J = \|\mathbf{z}\|^2 \mathbf{e}_1^\top \left(\log \mathbf{T}_{\mathbf{z}}^{(J)}\right) \mathbf{e}_1.$$

If $J < N$, CG underestimates the inverse quadratic term and overestimates the log determinant in expectation:

$$u_J \leq \mathbf{y}^\top \widehat{\mathbf{K}}_{\mathbf{XX}}^{-1}\mathbf{y}, \quad \mathbb{E}_{\mathbf{z}}[v_J] \geq \log|\widehat{\mathbf{K}}_{\mathbf{XX}}|. \tag{9}$$

The biases of both terms decay at a rate of $\mathcal{O}(C^{-2J})$, where C is a constant that depends on the conditioning of $\widehat{\mathbf{K}}_{\mathbf{XX}}$.



Thm 1: CG underfits the data

Loss = model complexity + fitting error

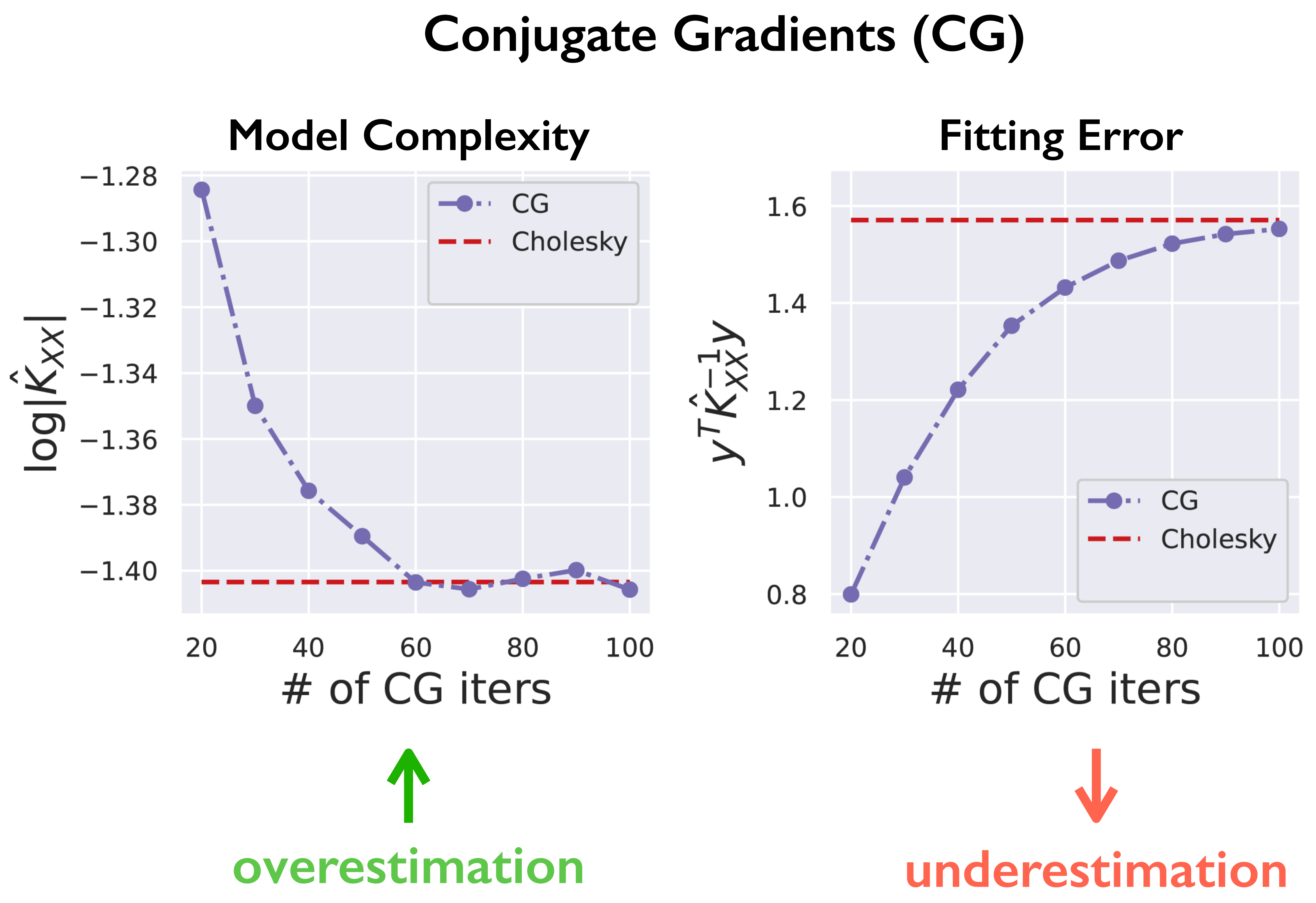
Theorem 1. *Let u_J and v_J be the estimates of $\mathbf{y}^\top \widehat{\mathbf{K}}_{\mathbf{XX}}^{-1} \mathbf{y}$ and $\log|\widehat{\mathbf{K}}_{\mathbf{XX}}|$ respectively after J iterations of CG; i.e.:*

$$u_J = \mathbf{y}^\top \left(\textstyle\sum_{i=1}^{J} \gamma_i \mathbf{d}_i \right), \quad v_J = \|\mathbf{z}\|^2 \mathbf{e}_1^\top \left(\log \mathbf{T}_{\mathbf{z}}^{(J)} \right) \mathbf{e}_1.$$

If $J < N$, CG underestimates the inverse quadratic term and overestimates the log determinant in expectation:

$$u_J \leq \mathbf{y}^\top \widehat{\mathbf{K}}_{\mathbf{XX}}^{-1} \mathbf{y}, \quad \mathbb{E}_{\mathbf{z}}[v_J] \geq \log|\widehat{\mathbf{K}}_{\mathbf{XX}}|. \qquad (9)$$

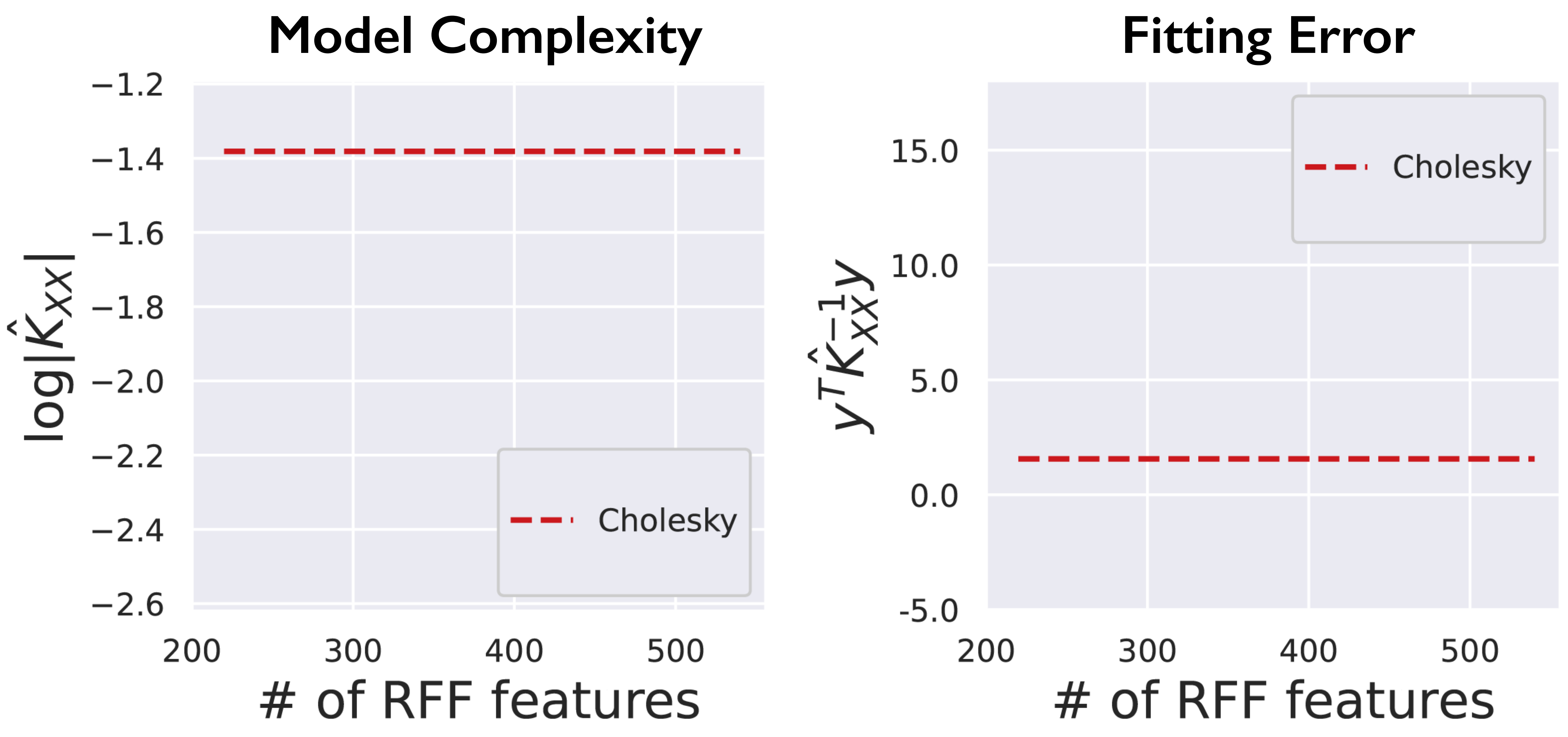
The biases of both terms decay at a rate of $\mathcal{O}(C^{-2J})$, where C is a constant that depends on the conditioning of $\widehat{\mathbf{K}}_{\mathbf{XX}}$.



Thm 2: RFF overfits the data

Loss = model complexity + fitting error

Random Fourier Features (RFF)



Theorem 2. *Let $\widetilde{\mathbf{K}}_J$ be the RFF approximation with $J/2$ random features. In expectation, $\widetilde{\mathbf{K}}_J$ overestimates the inverse quadratic and underestimates the log determinant:*

$$\mathbb{E}_{\mathbb{P}(\boldsymbol{\omega})}\left[\mathbf{y}^\top \widetilde{\mathbf{K}}_J^{-1}\mathbf{y}\right] \geq \mathbf{y}^\top \widehat{\mathbf{K}}_{\mathbf{XX}}^{-1}\mathbf{y} \qquad (10)$$

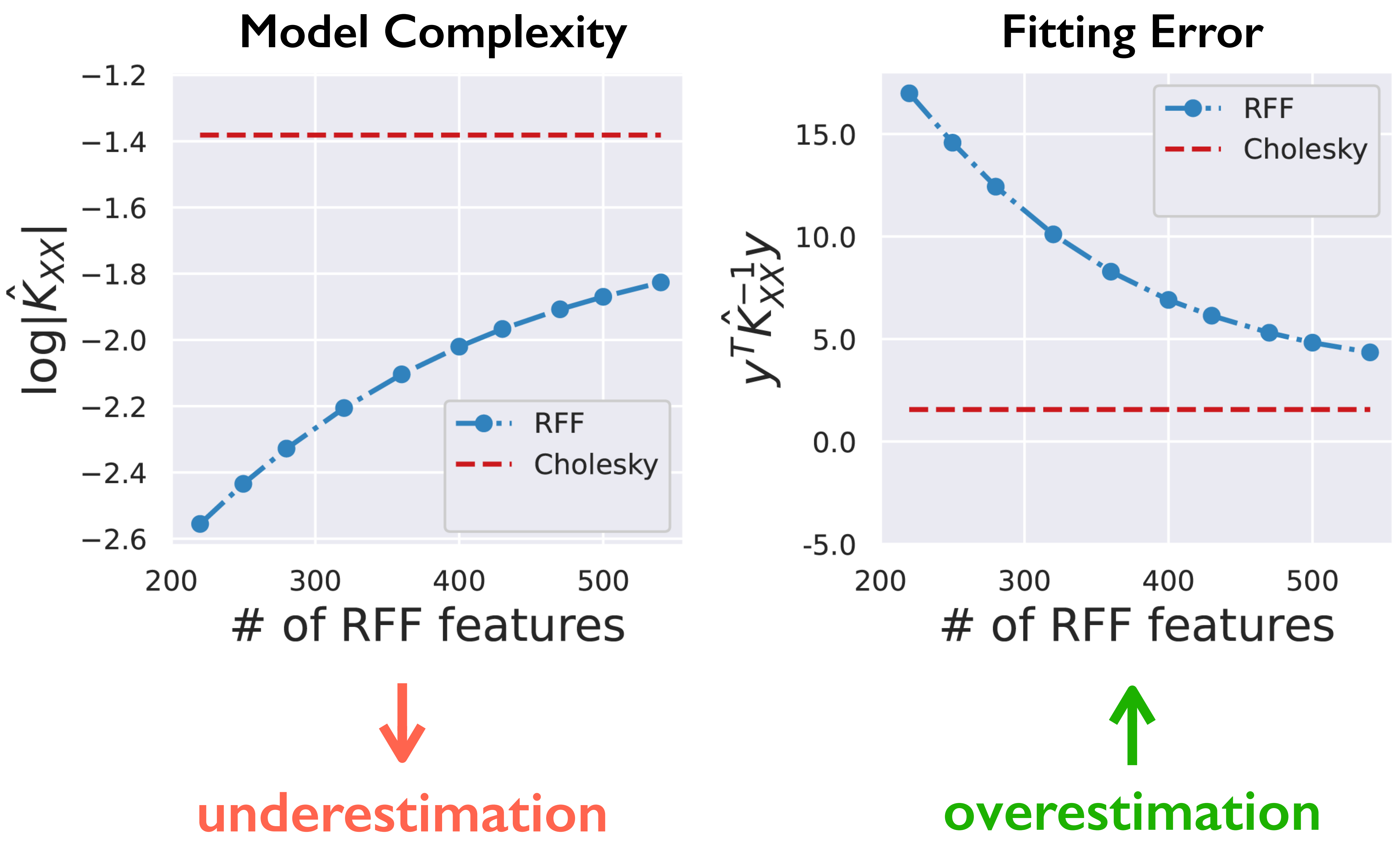
$$\mathbb{E}_{\mathbb{P}(\boldsymbol{\omega})}\left[\log|\widetilde{\mathbf{K}}_J|\right] \leq \log|\widehat{\mathbf{K}}_{\mathbf{XX}}|. \qquad (11)$$

The biases of both terms decay at a rate of $\mathcal{O}(1/J)$.

Thm 2: RFF overfits the data

Loss = model complexity + fitting error

Random Fourier Features (RFF)



Theorem 2. *Let $\widetilde{\mathbf{K}}_J$ be the RFF approximation with $J/2$ random features. In expectation, $\widetilde{\mathbf{K}}_J$ overestimates the inverse quadratic and underestimates the log determinant:*

$$\mathbb{E}_{\mathbb{P}(\boldsymbol{\omega})}\left[\mathbf{y}^\top \widetilde{\mathbf{K}}_J^{-1}\mathbf{y}\right] \geq \mathbf{y}^\top \widehat{\mathbf{K}}_{\mathbf{XX}}^{-1}\mathbf{y} \tag{10}$$

$$\mathbb{E}_{\mathbb{P}(\boldsymbol{\omega})}\left[\log|\widetilde{\mathbf{K}}_J|\right] \leq \log|\widehat{\mathbf{K}}_{\mathbf{XX}}|. \tag{11}$$

The biases of both terms decay at a rate of $\mathcal{O}(1/J)$.

Our method: Bias elimination via randomized truncation

computation VS ~~bias~~ variance

Single Sample RFF

Lynne et al., 2015, Beatson & Adams, 2018

Random Fourier Features (RFF)

$$\mathbf{K} \approx \sum_{j=1}^{J} \phi_j \phi_j^\top$$

Conjugate Gradients (CG)

$$\mathbf{K}^{-1}\mathbf{y} \approx \sum_{j=1}^{J} \gamma_j \mathbf{d}_j$$

Russian Roulette CG

Kahn, 1955, Beatson & Adams, 2018, Chen et al., 2019

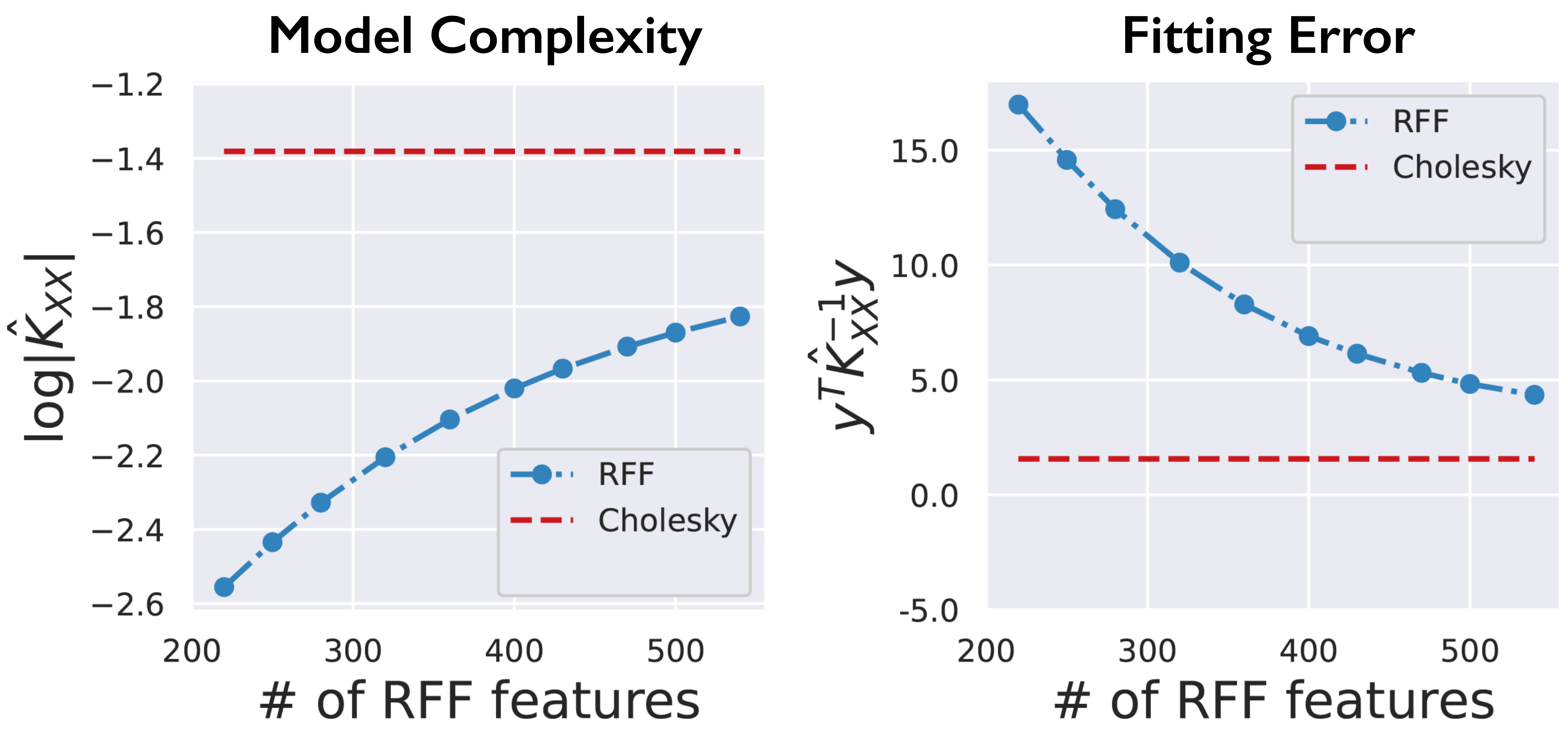
$$J \sim$$

(& re-weight each iteration)

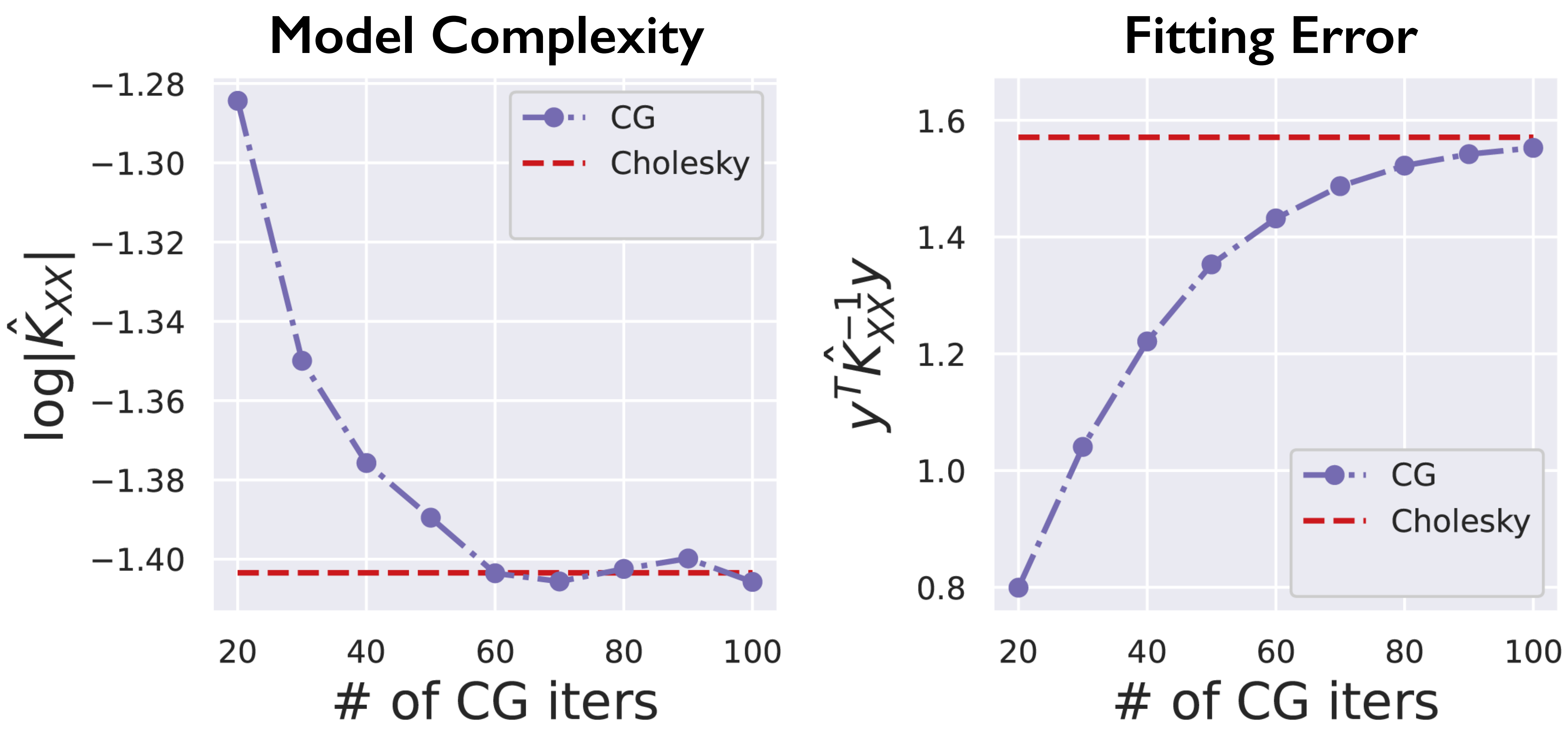
Our method: Bias elimination via randomized truncation

Loss = model complexity + fitting error

Random Fourier Features (RFF)



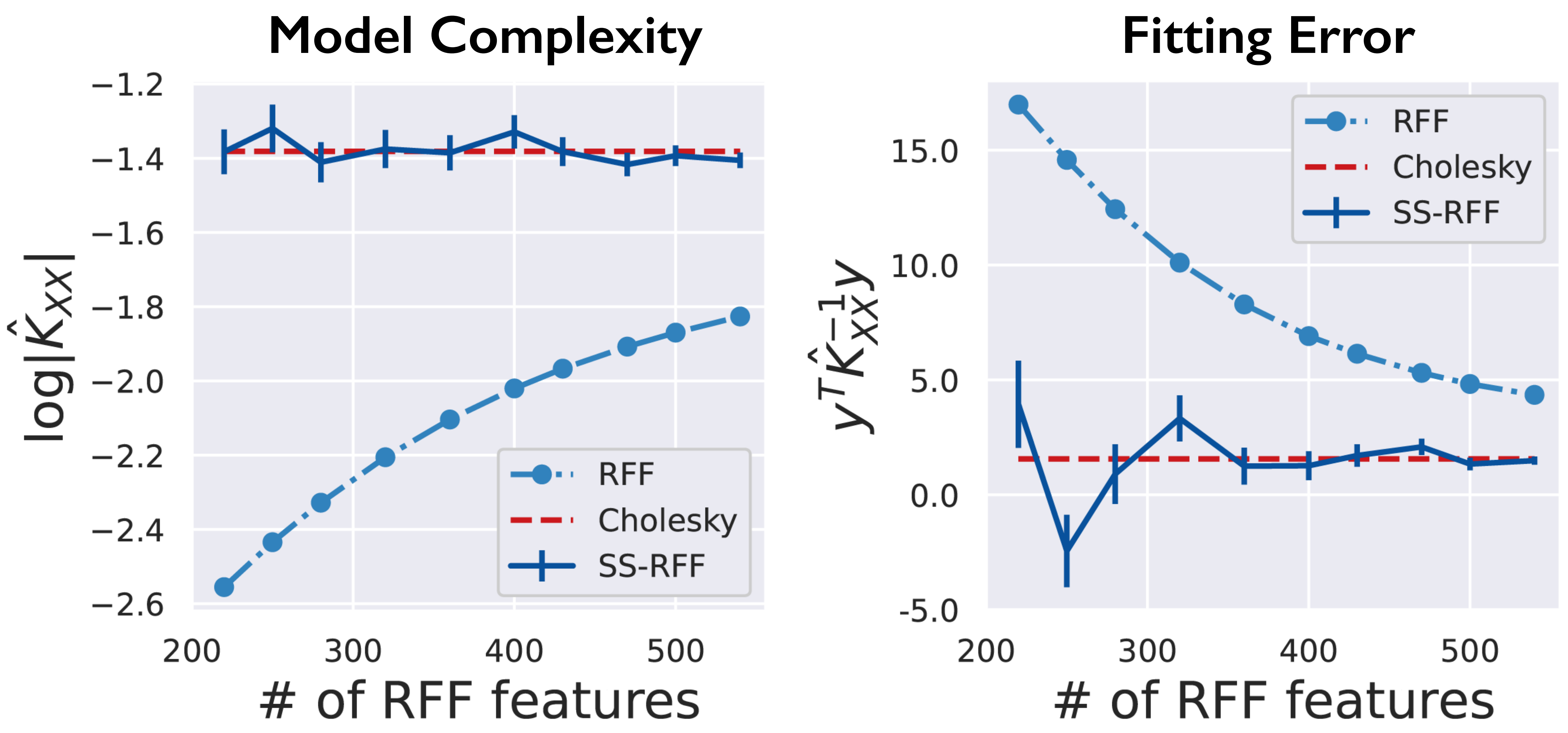
Conjugate Gradients (CG)



Our method: Bias elimination via randomized truncation

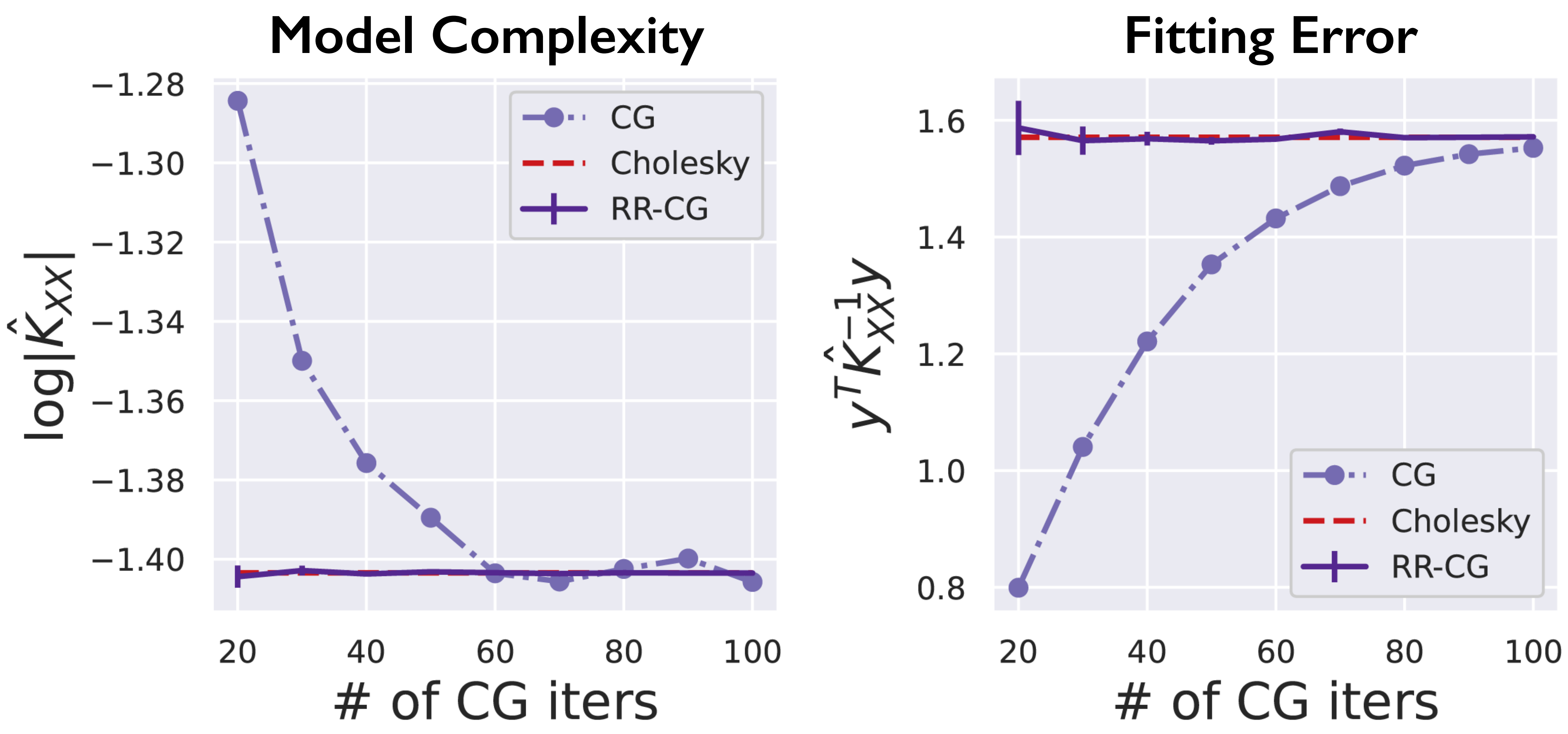
Loss = model complexity + fitting error

Random Fourier Features (RFF)



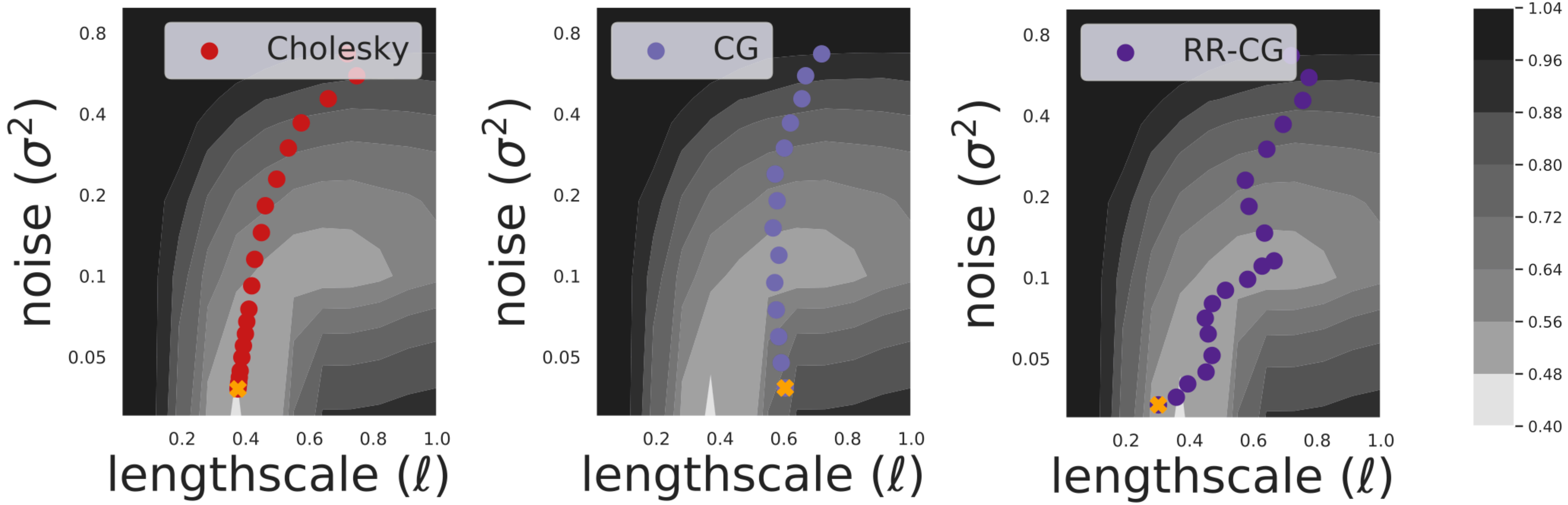
Single Sample RFF

Conjugate Gradients (CG)



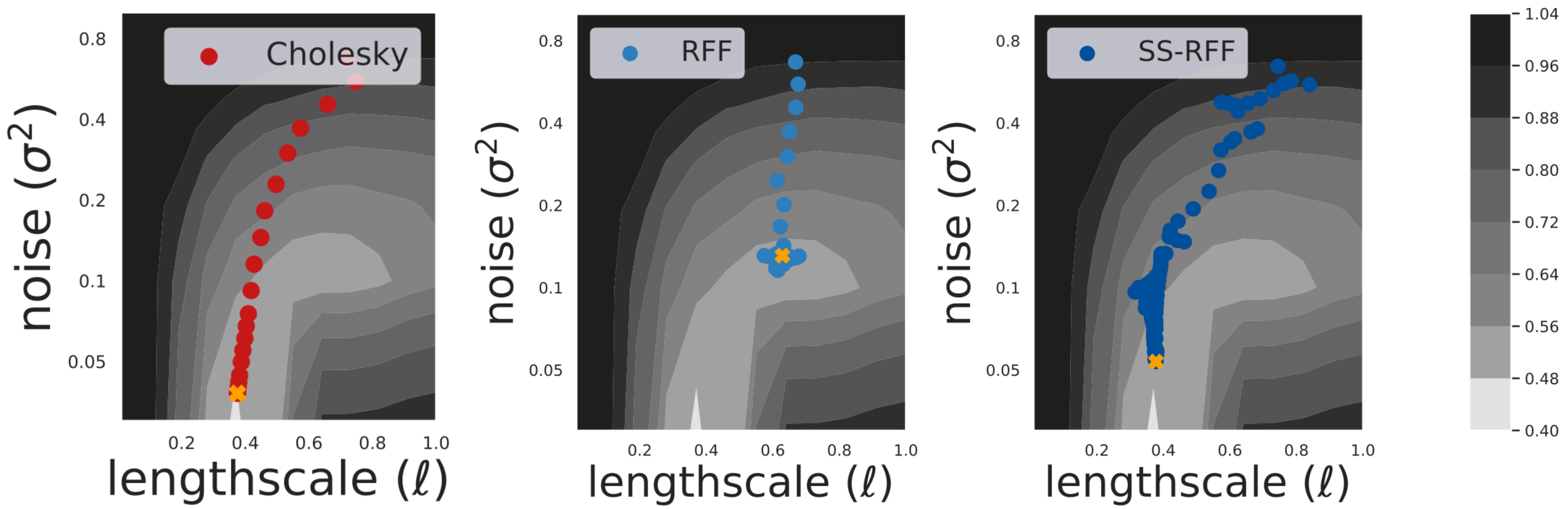
Russian Roulette CG

RR-CG achieves superior loss values



state-of-the-art performance on large-scale datasets with $\mathcal{O}(N^2)$ computation

SS-RFF achieves superior loss values



slow convergence on large-scale datasets due to auxiliary variance

What did we discover?

proving systematic biases in scalable GPs

novel method to eliminate the biases via randomized truncations

https://github.com/cunningham-lab/RTGPS