Whitening and Second Order Optimization Both Make Information in the Dataset Unusable During Training, and Can Reduce or Prevent Generalization

Neha S. Wadia,[1] Daniel Duckworth,[2] Samuel S. Schoenholz,[2]
Ethan Dyer,[2] Jascha Sohl-Dickstein[2]

[1]University of California, Berkeley (Work done as an intern at Google Brain.)

[2]Google Brain

ICML 2021

Theoretical Results

Result 1. For models with a dense, isotropically initialized first layer, the only information SGD can use to generalize is contained in the sample second moment matrix (Gram matrix).

Theoretical Results

Result 1. For models with a dense, isotropically initialized first layer, the only information SGD can use to generalize is contained in the sample second moment matrix (Gram matrix).

Result 2. Whitening removes information in this matrix in a dimensionality-dependent manner.

Theoretical Results

Result 1. For models with a dense, isotropically initialized first layer, the only information SGD can use to generalize is contained in the sample second moment matrix (Gram matrix).

Result 2. Whitening removes information in this matrix in a dimensionality-dependent manner.

$\Rightarrow$ Whitening negatively impacts generalization in a dimensionality-dependent manner.

Theoretical Results

Result 1. For models with a dense, isotropically initialized first layer, the only information SGD can use to generalize is contained in the sample second moment matrix (Gram matrix).

Result 2. Whitening removes information in this matrix in a dimensionality-dependent manner.

$\Rightarrow$ Whitening negatively impacts generalization in a dimensionality-dependent manner.

Result 3. $\exists$ an equivalence between Newton's method on unwhitened data and SGD on whitened data in linear models and overparametrized networks.

Theoretical Results

Result 1. For models with a dense, isotropically initialized first layer, the only information SGD can use to generalize is contained in the sample second moment matrix (Gram matrix).

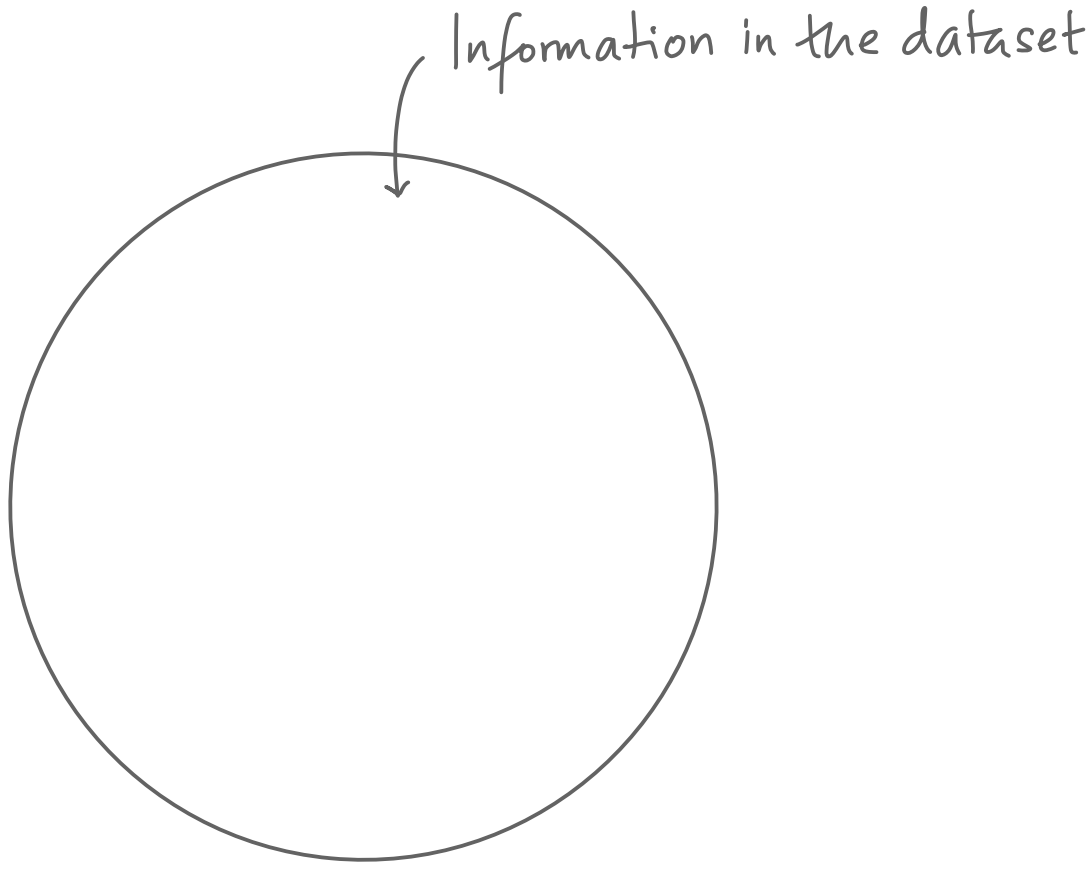
Result 2. Whitening removes information in this matrix in a dimensionality-dependent manner.

$\Rightarrow$ Whitening negatively impacts generalization in a dimensionality-dependent manner.

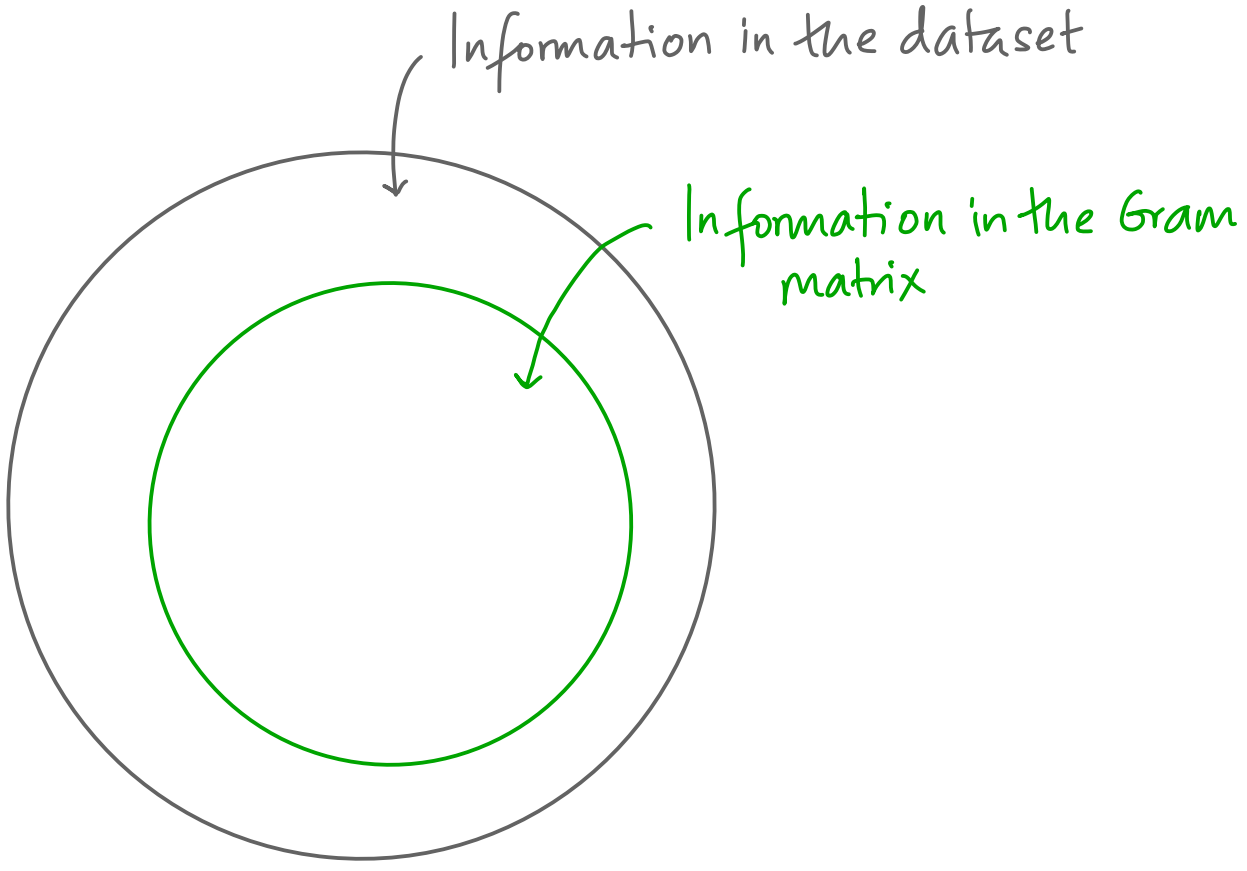
Result 3. $\exists$ an equivalence between Newton's method on unwhitened data and SGD on whitened data in linear models and overparametrized networks.

$\Rightarrow$ Generalization in these models is similarly harmed.

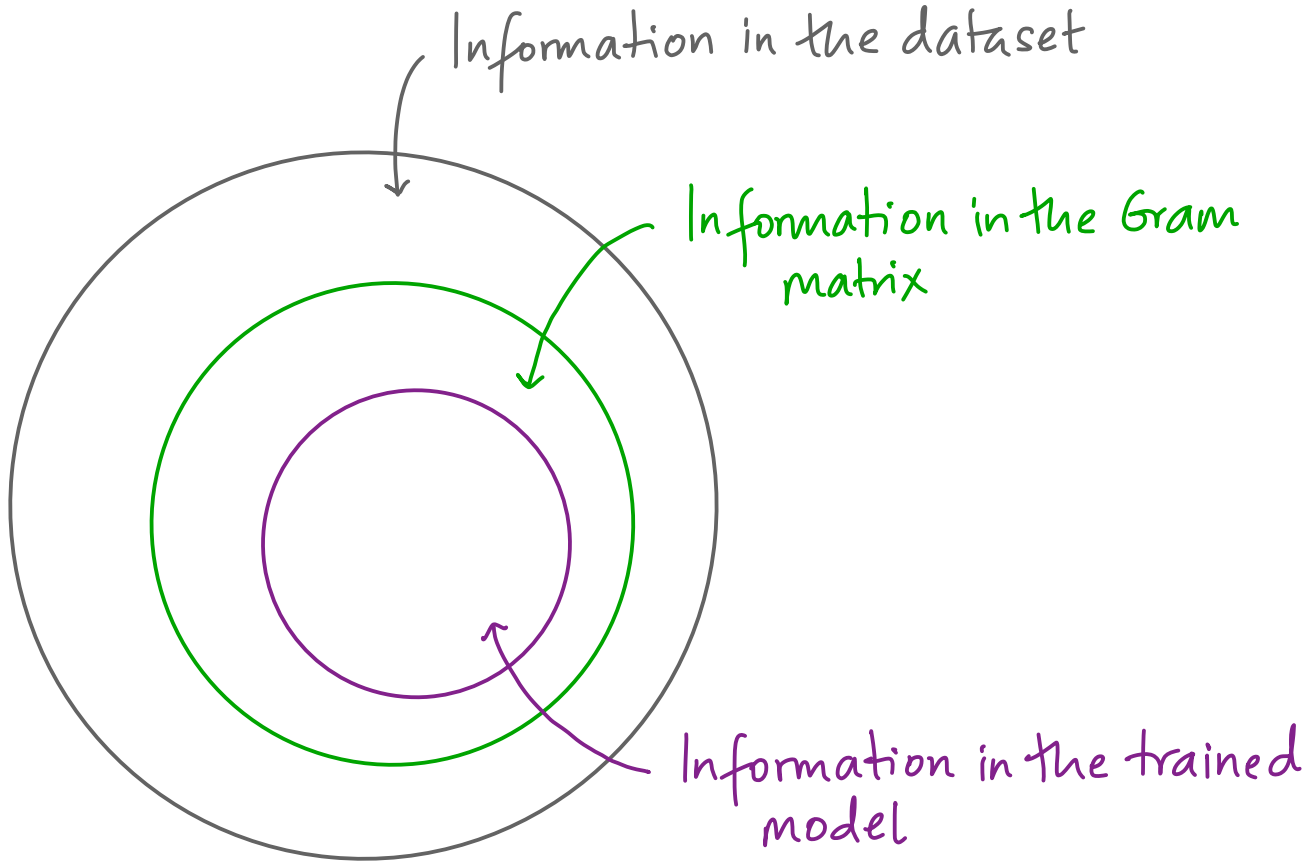
Theoretical Results



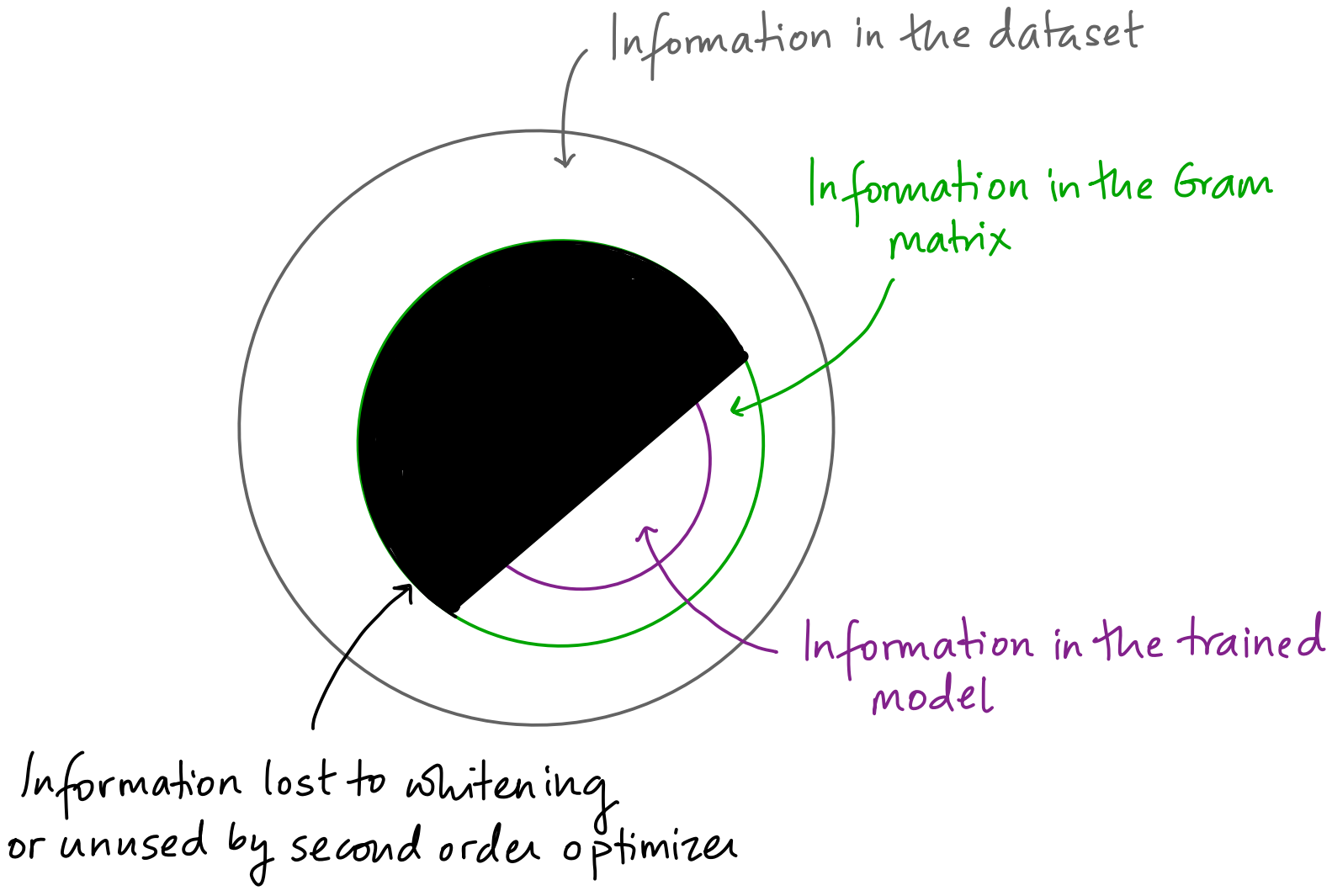
Theoretical Results



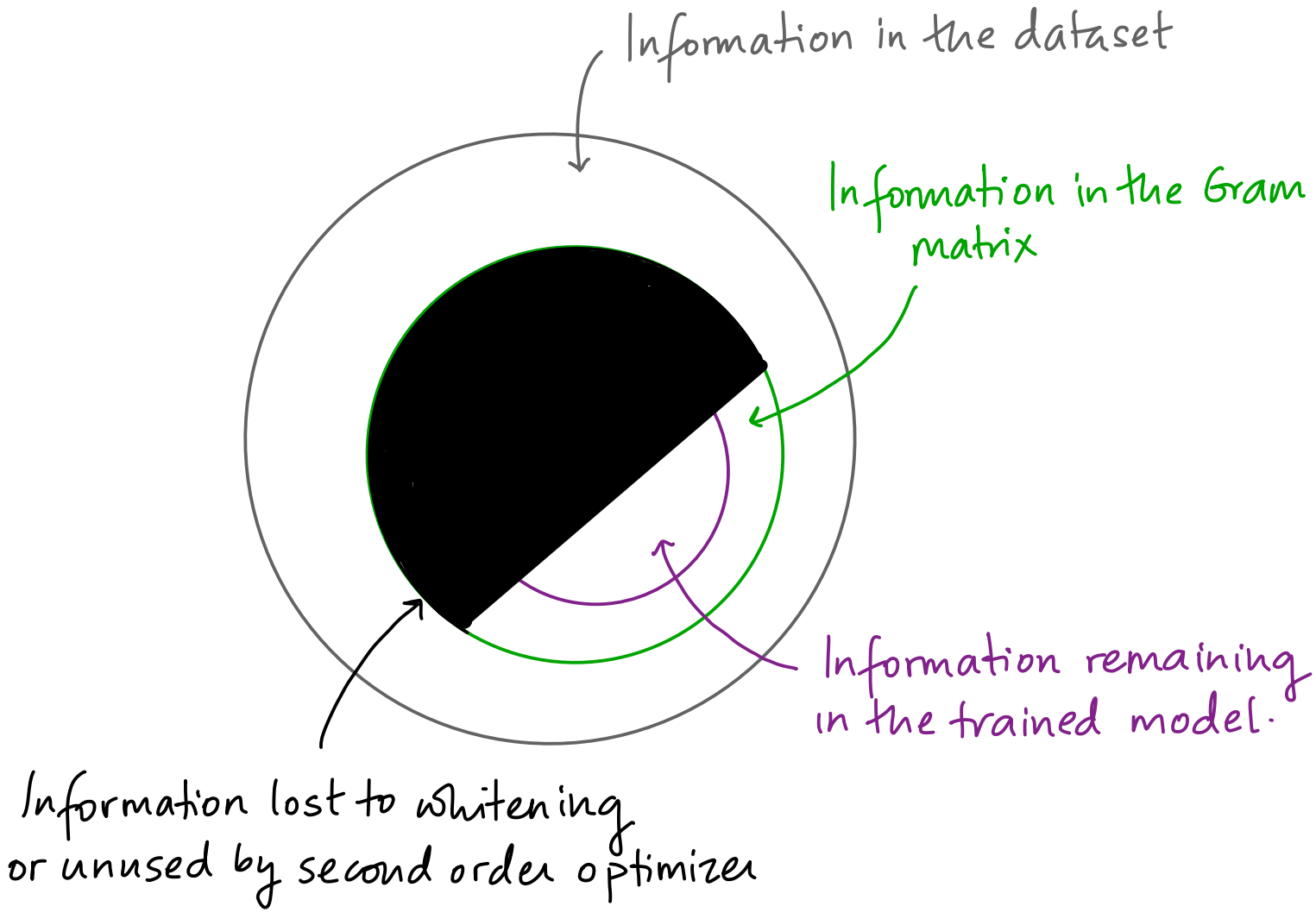
Theoretical Results



Theoretical Results



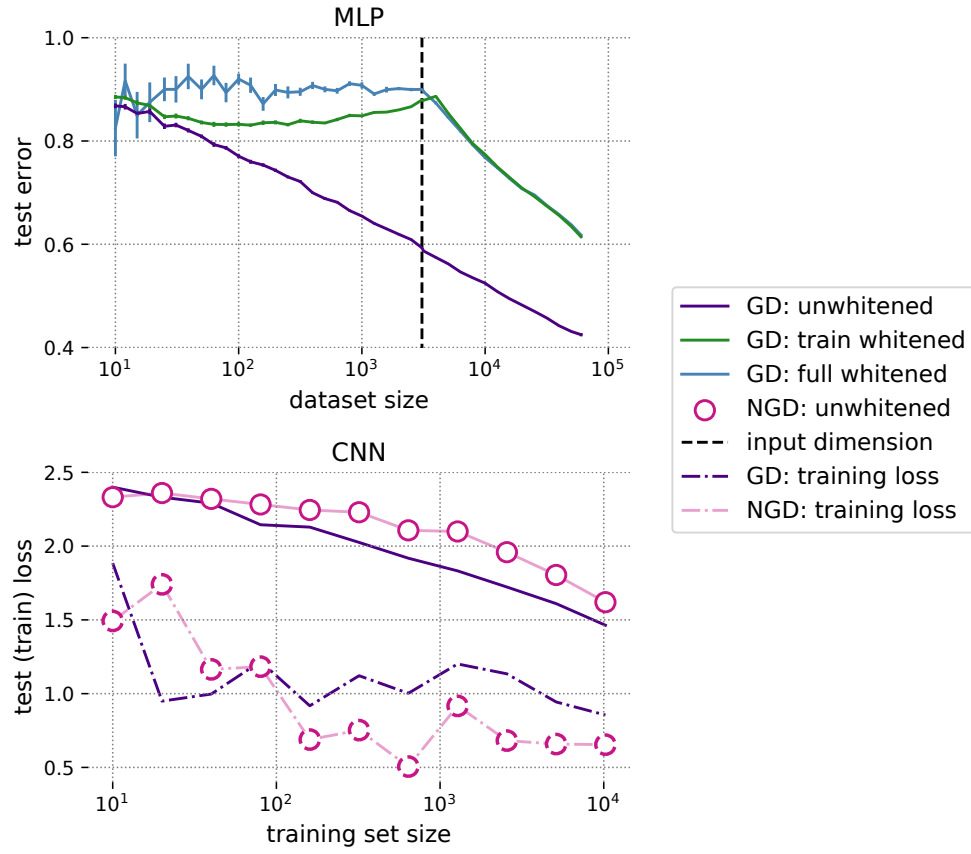
Theoretical Results



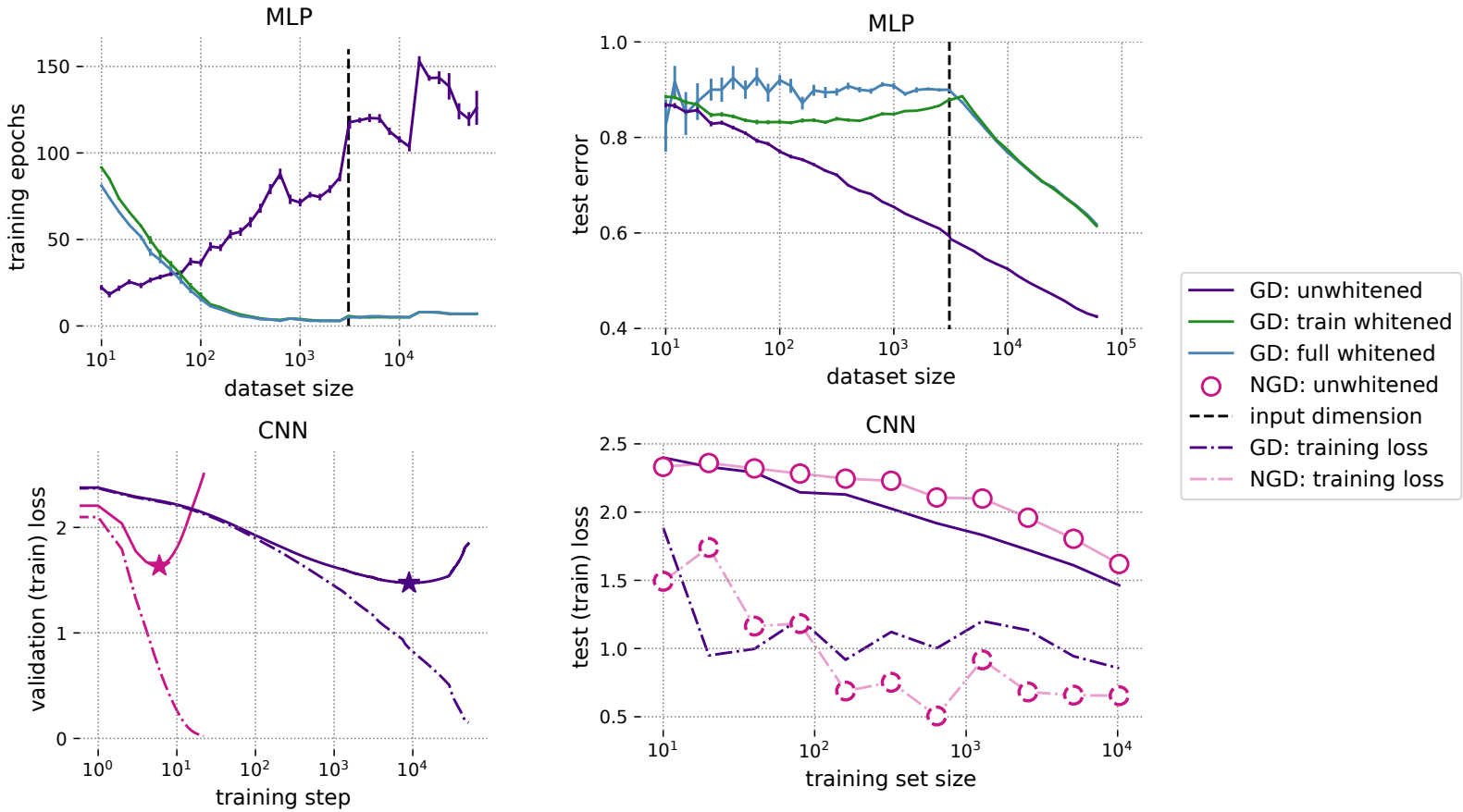
Main message

Whitening always throws away information; pure second order optimizers fail to use that information.

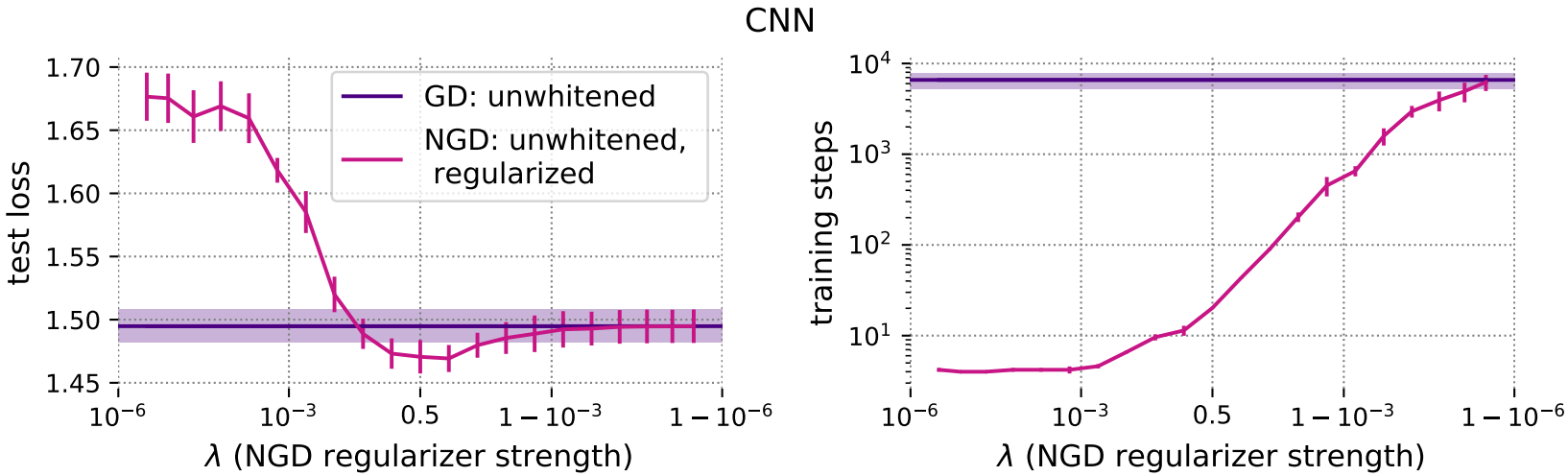
Whitening and pure second order optimization harm generalization



Whitening and pure second order optimization harm generalization, but both speed up training



Regularized second order optimization can sometimes train faster *and* generalize better



NGD preconditioner: $((1-\lambda)B + \lambda I)^{-1}$, $\lambda \in [0, 1]$, B: Hessian.

Summary

- Whitening and pure second order optimization both cause a reduction in generalization through an information loss mechanism, ...
- ... but require fewer iterations to train.
- Regularized second order optimizers can in some cases both train faster and generalize better than SGD.

Thank you!
Please come to our poster and check out our paper!