Debiasing Mini-Batch Quadratics for Applications in Deep Learning

Lukas Tatzel, Bálint Mucsányi, Osane Hackel & Philipp Hennig

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Max Planck Institute for Intelligent Systems

DFG Deutsche Forschungsgemeinschaft

Tübingen AI Center

imprs-is

erc some of the presented work is supported by the European Research Council.

Why Are Mini-Batch Quadratics Important?

Notation.

- Regularized loss: $\mathcal{L}_{\text{reg}}(\boldsymbol{\theta}; \mathcal{D}) = \frac{1}{|\mathcal{D}|} \sum_{n \in \mathcal{D}} \ell(f_{\boldsymbol{\theta}}(\mathbf{x}_n), \mathbf{y}_n) + r(\boldsymbol{\theta})$
- *Full-batch* quadratic around $\boldsymbol{\theta}_0$:

$$\mathcal{L}_{\text{reg}}(\boldsymbol{\theta}; \mathcal{D}) \approx q(\theta; \mathcal{D}) = \frac{1}{2}(\theta - \theta_0)^\top \mathbf{H}_{\mathcal{D}}(\theta - \theta_0) + (\theta - \theta_0)^\top \mathbf{g}_{\mathcal{D}} + c_{\mathcal{D}}$$

Why Are Mini-Batch Quadratics Important?

Notation.

- Regularized loss: $\mathcal{L}_{\text{reg}}(\boldsymbol{\theta}; \mathcal{D}) = \frac{1}{|\mathcal{D}|} \sum_{n \in \mathcal{D}} \ell(f_{\boldsymbol{\theta}}(\mathbf{x}_n), \mathbf{y}_n) + r(\boldsymbol{\theta})$
- *Full-batch* quadratic around $\boldsymbol{\theta}_0$:

$$\mathcal{L}_{\text{reg}}(\boldsymbol{\theta}; \mathcal{D}) \approx q(\boldsymbol{\theta}; \mathcal{D}) = \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)^\top \mathbf{H}_{\mathcal{D}}(\boldsymbol{\theta} - \boldsymbol{\theta}_0) + (\boldsymbol{\theta} - \boldsymbol{\theta}_0)^\top \mathbf{g}_{\mathcal{D}} + c_{\mathcal{D}}$$

- *Mini-batch* quadratic around $\boldsymbol{\theta}_0$: $\mathcal{L}_{\text{reg}}(\boldsymbol{\theta}; \mathcal{D}) \approx q(\boldsymbol{\theta}; \mathcal{D}) \approx q(\boldsymbol{\theta}; \mathcal{B})$ with $\mathcal{B} \subset \mathcal{D}$

Why Are Mini-Batch Quadratics Important?

Notation.

- Regularized loss: $\mathcal{L}_{\text{reg}}(\boldsymbol{\theta}; \mathcal{D}) = \frac{1}{|\mathcal{D}|} \sum_{n \in \mathcal{D}} \ell(f_{\boldsymbol{\theta}}(\mathbf{x}_n), \mathbf{y}_n) + r(\boldsymbol{\theta})$
- *Full-batch* quadratic around $\boldsymbol{\theta}_0$:

$$\mathcal{L}_{\text{reg}}(\boldsymbol{\theta}; \mathcal{D}) \approx q(\theta; \mathcal{D}) = \frac{1}{2}(\theta - \theta_0)^\top \mathbf{H}_{\mathcal{D}}(\theta - \theta_0) + (\theta - \theta_0)^\top \mathbf{g}_{\mathcal{D}} + c_{\mathcal{D}}$$

- *Mini-batch* quadratic around $\boldsymbol{\theta}_0$: $\mathcal{L}_{\text{reg}}(\boldsymbol{\theta}; \mathcal{D}) \approx q(\theta; \mathcal{D}) \approx q(\theta; \mathcal{B})$ with $\mathcal{B} \subset \mathcal{D}$

What are these quadratic approximations used for?

- Second-order optimizers rely on Newton steps: $\boldsymbol{\theta}_{\text{new}} = \arg\min_{\boldsymbol{\theta}} q(\theta; \mathcal{B})$
- Post-hoc uncertainty quantification with the Laplace approximation: $q(\theta; \mathcal{B}) \rightsquigarrow \mathcal{N}(\boldsymbol{\theta}; \boldsymbol{\theta}_\star, \mathbf{H}_{\mathcal{B}}^{-1})$

Why Are Mini-Batch Quadratics Important?

Notation.

- Regularized loss: $\mathcal{L}_{\text{reg}}(\boldsymbol{\theta}; \mathcal{D}) = \frac{1}{|\mathcal{D}|} \sum_{n \in \mathcal{D}} \ell(f_{\boldsymbol{\theta}}(\mathbf{x}_n), \mathbf{y}_n) + r(\boldsymbol{\theta})$
- *Full-batch* quadratic around $\boldsymbol{\theta}_0$:

$$\mathcal{L}_{\text{reg}}(\boldsymbol{\theta}; \mathcal{D}) \approx q(\boldsymbol{\theta}; \mathcal{D}) = \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)^\top \mathbf{H}_{\mathcal{D}}(\boldsymbol{\theta} - \boldsymbol{\theta}_0) + (\boldsymbol{\theta} - \boldsymbol{\theta}_0)^\top \mathbf{g}_{\mathcal{D}} + c_{\mathcal{D}}$$

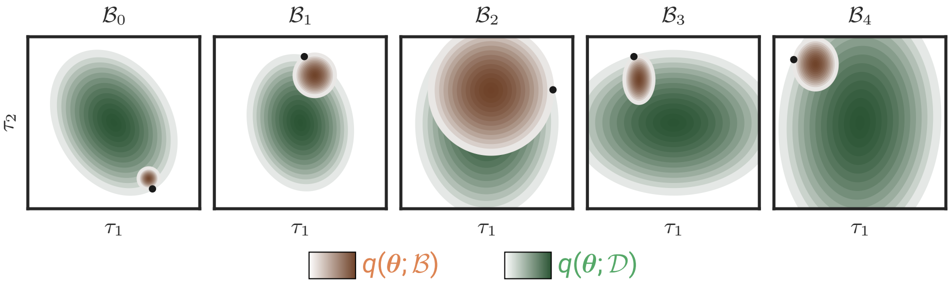
- *Mini-batch* quadratic around $\boldsymbol{\theta}_0$: $\mathcal{L}_{\text{reg}}(\boldsymbol{\theta}; \mathcal{D}) \approx q(\boldsymbol{\theta}; \mathcal{D}) \approx q(\boldsymbol{\theta}; \mathcal{B})$ with $\mathcal{B} \subset \mathcal{D}$

What are these quadratic approximations used for?

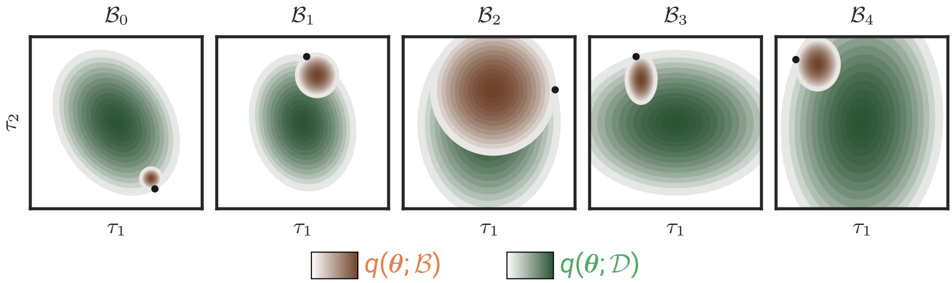
- Second-order optimizers rely on Newton steps: $\boldsymbol{\theta}_{\text{new}} = \arg\min_{\boldsymbol{\theta}} q(\boldsymbol{\theta}; \mathcal{B})$
- Post-hoc uncertainty quantification with the Laplace approximation: $q(\boldsymbol{\theta}; \mathcal{B}) \rightsquigarrow \mathcal{N}(\boldsymbol{\theta}; \boldsymbol{\theta}_\star, \mathbf{H}_{\mathcal{B}}^{-1})$

Is $q(\boldsymbol{\theta}; \mathcal{B})$ a good proxy for $q(\boldsymbol{\theta}; \mathcal{D})$?

Mini-Batch Quadratics Are Biased!

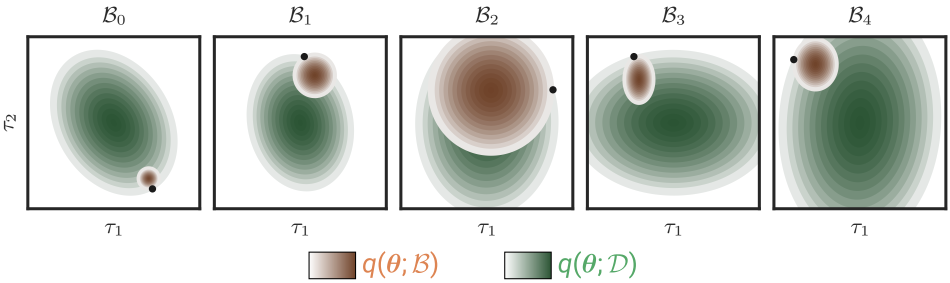


Mini-Batch Quadratics Are Biased!



Quadratic approximations to the training loss computed on *mini-batches* of the training data provide a *distorted* representation of the true underlying loss landscape.

Mini-Batch Quadratics Are Biased!



Quadratic approximations to the training loss computed on *mini-batches* of the training data provide a *distorted* representation of the true underlying loss landscape.

This bias is highly relevant for applications that operate on $q(\theta; \mathcal{B})$.

- Second-order optimization: Detrimental updates of the parameters
- Laplace approximation: Unreliable uncertainty estimates

Can We Fix It?

Idea: Split mini-batch in two, use one half for directions, the other for directional quantities.

We can eliminate the bias at almost no computational overhead by splitting the mini-batch in two!

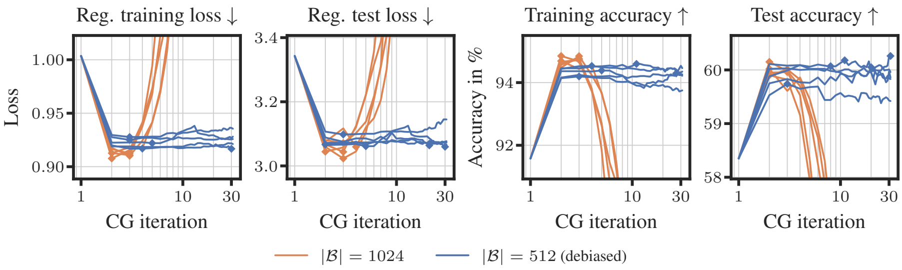
Can We Fix It?

Idea: Split mini-batch in two, use one half for directions, the other for directional quantities.

We can eliminate the bias at almost no computational overhead by splitting the mini-batch in two!

Empirical results.

- Debiased CG is much more stable than the single-batch approach.



- Debiased Laplace approximation mimics the full-batch Laplace approximation very well.

Summary

Contributions.

1. We show that the bias presented here introduces a systematic error.
2. We provide a theoretical explanation.
3. We explain the relevance of this bias for second-order optimization and uncertainty quantification via the Laplace approximation in deep learning.
4. We develop debiasing strategies and demonstrate their effectiveness.

The paper can be found at: `https://openreview.net/forum?id=Q0TEVKV2cp`

Thank You for Your Attention!