Adaptive Step Sizes for Stochastic Gradient Descent

Stochastic Optimization Problems

Consider a family of functions, indexed with $\xi \in \Omega$, for some probability space $(\Omega, \mathbb{P})$:

$$f_\xi : \mathbb{R}^n \to \mathbb{R}$$

Stochastic Optimization aims at minimizing:

$$F = x \mapsto \mathbb{E}_\xi \left[f_\xi(x) \right] = \int_\Omega f_\xi(x) \mathrm{d}\mathbb{P}(\xi) \qquad (1)$$

where $\mathbb{P}$ is the probability measure on Ω.

Gradient Descent

A simple way to solve smooth optimization problems like

$$\min_{x \in \mathbb{R}^n} F(x)$$

is **gradient descent**:

- Select initial $x_0 \in \mathbb{R}^n$
- In Iteration k, compute $\nabla F(x_k)$ and update
$$x_{k+1} = x_k - \alpha_k \nabla f(x_k).$$
- Here: $\alpha_k > 0$ is some **step size**.

Practical Approach: (Mini-)Batch-Algorithms

- Usually, the measure $\mathbb{P}$ is not available. Instead, one has access to a set of *observations* $\xi_1, \dots, \xi_N \overset{\text{iid}}{\sim} \mathbb{P}$.
- Then
$$f_\xi(x) = \frac{1}{n} \sum_{i=1}^{n} f_{\xi_i}(x)$$
is a (stochastic) approximation to F from (1).
- In each iteration, sample a new ξ_k and use $\nabla f_{\xi_k}(x_k)$ as a search direction.
- This search direction is an **unbiased** estimator for $\nabla F(x)$:
$$\mathbb{E}_\xi \left[\nabla f_\xi \right] = \nabla F(x).$$
- This algorithm is called **Stochastic Gradient Descent** (SGD).

Numerical Results on Artificial Data

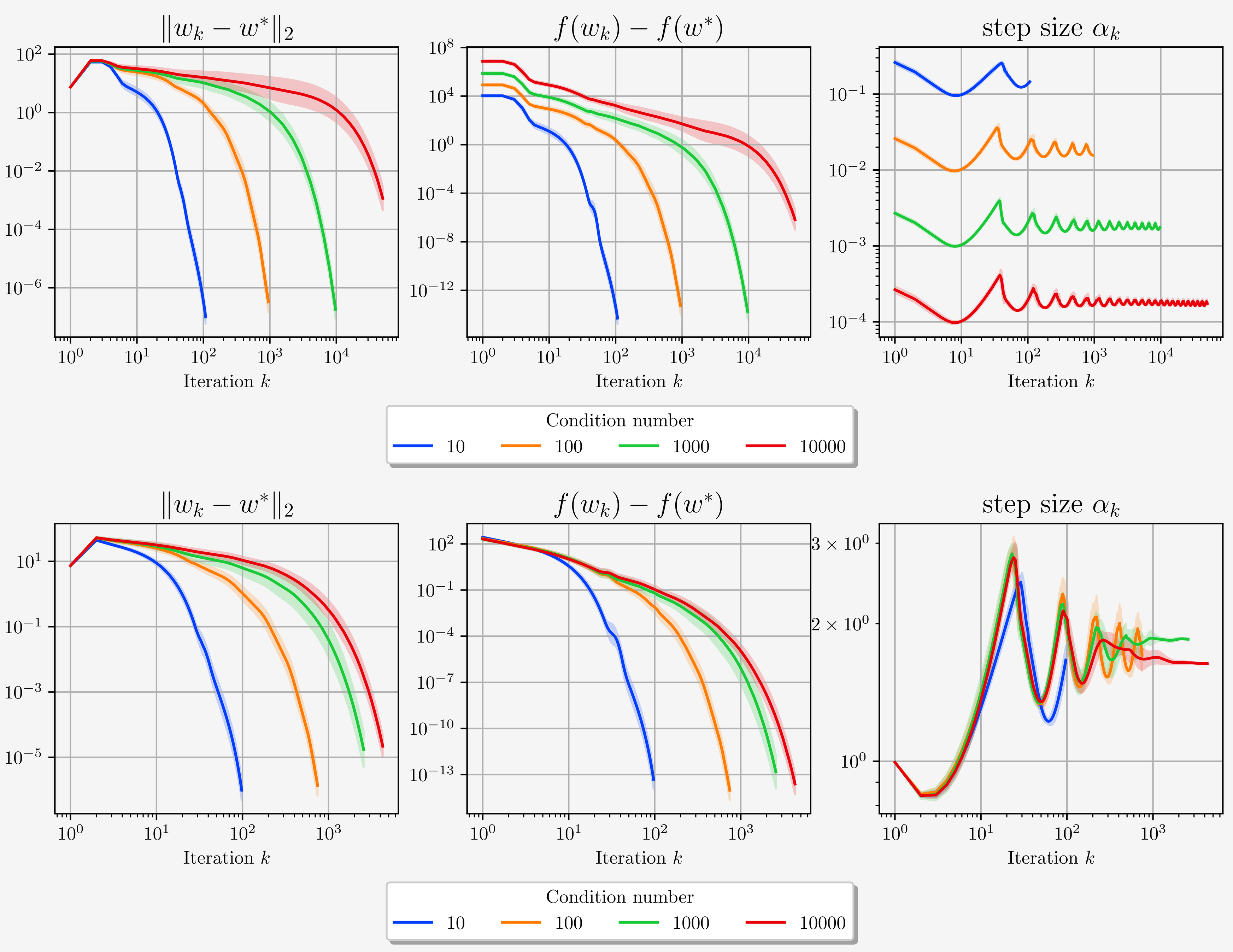

Figure: Performance on *interpolating* artificial data. For non-interpolating see [1]

Convergence Analysis & Variance

- Main theoretical concern: Noise in the search direction:
$$\mathbb{V}_\xi \left[\nabla f_\xi(x_k) \right] = \mathbb{E}_\xi \left[\| \nabla f_\xi(x_k) - \nabla F(x_k) \|^2 \right].$$
- Variance models are needed. A popular choice is:
$$\mathbb{V}_\xi \left[\nabla f_\xi(x_k) \right] \le V_0 + V_1 \| \nabla F(x_k) \|^2 \qquad (2)$$
- Such models can be used for step size control and a-priori error analysis.
- In fact, bounds as in (2) can be deduced form certain smoothness- and convexity-assumptions ([1, 2]).
- However, they might lead to unwanted dependency of the step size on the convexity.
- In [1] we developed an alternative model, which mitigates these problems.

Direct Incoporation of the Variance

- An Alternative Approach directly uses the variance for step size selection.
- In general, it holds that
$$\alpha_k = \frac{\mathbb{E}_\xi \left[\| \nabla f_\xi(w_k) \|^2 \right] - \mathbb{V}_\xi [\nabla f_\xi(w_k)]}{L \, \mathbb{E}_\xi \left[\| \nabla f_\xi(w_k) \|^2 \right]} \qquad (3)$$
is the step size which maximizes the *expected descent in the current iteration*.
- In [1] we developed techniques to estimate the parameters needed in (3).
- We use exponential smoothing techniques to minimize noise in the observed quantities.
- Moderate computational overhead leads to a **nearly hyperparameter free** stochastic optimization algorithm.
- Theoretical convergence guarantees are given in special cases, numerical experiments show the methods works well beyond theory.

Numerical Results on Image Classification Tasks

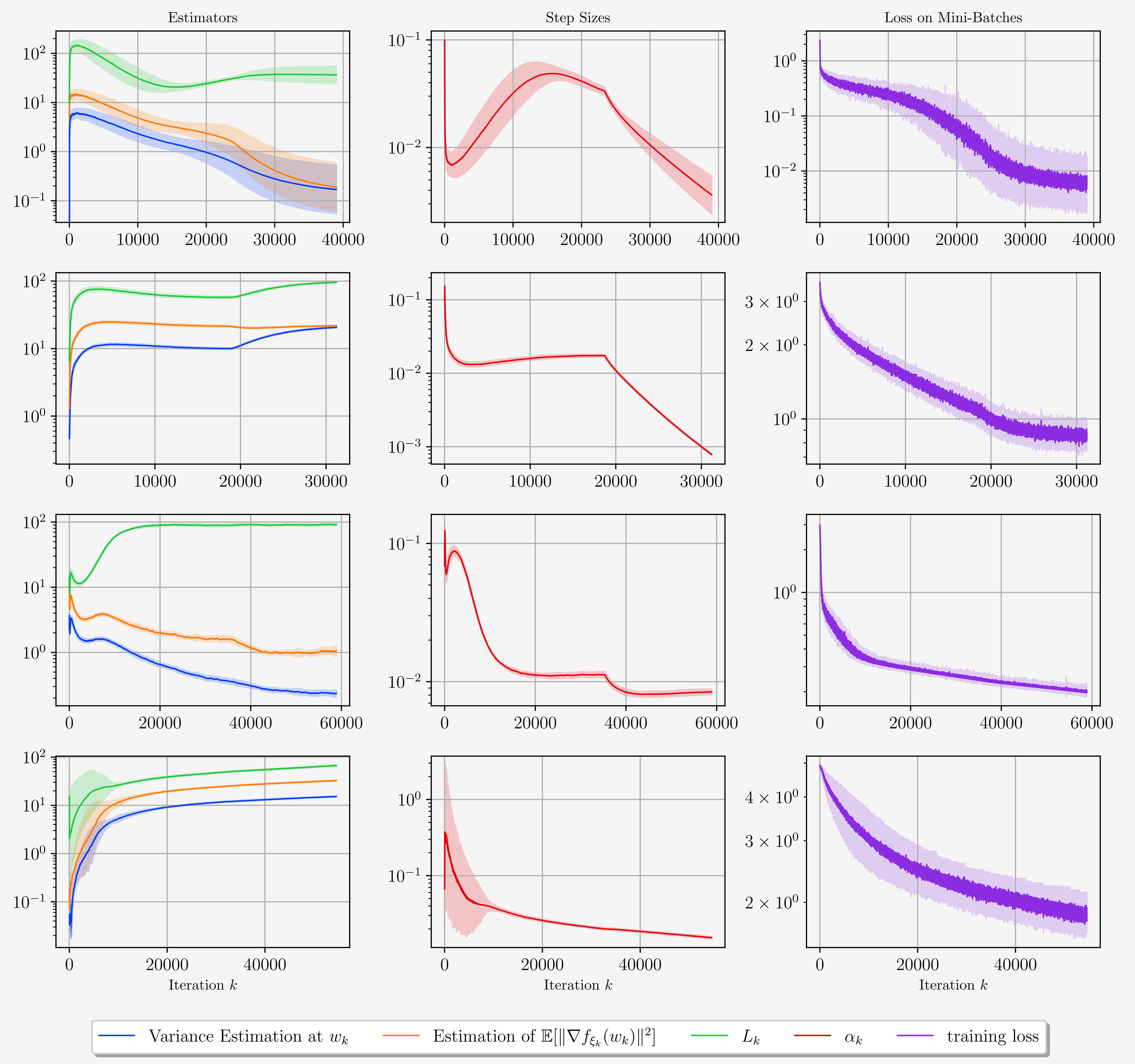

Figure: Performance on Image Classification data sets. Top row: Fashion-MNIST, second row: CIFAR-10, third row: SVHN, last row: CIFAR-100.

References

[1] F. Köhne; L. Kreis; A. Schiela; R. Herzog. *Adaptive step sizes for preconditioned stochastic gradient descent.* 2023. arXiv: 2311.16956.

[2] L. M. Nguyen et al. *SGD and Hogwild! Convergence without the bounded gradients assumption.* 2018. arXiv: 1802.03801.

Our Paper

Frederik Köhne, Anton Schiela

frederik.koehne@uni-bayreuth.de, anton.schiela@uni-bayreuth.de

Chair of Applied Mathematics, Institute of Mathematics

Funded by DFG Deutsche Forschungsgemeinschaft German Research Foundation

UNIVERSITÄT BAYREUTH