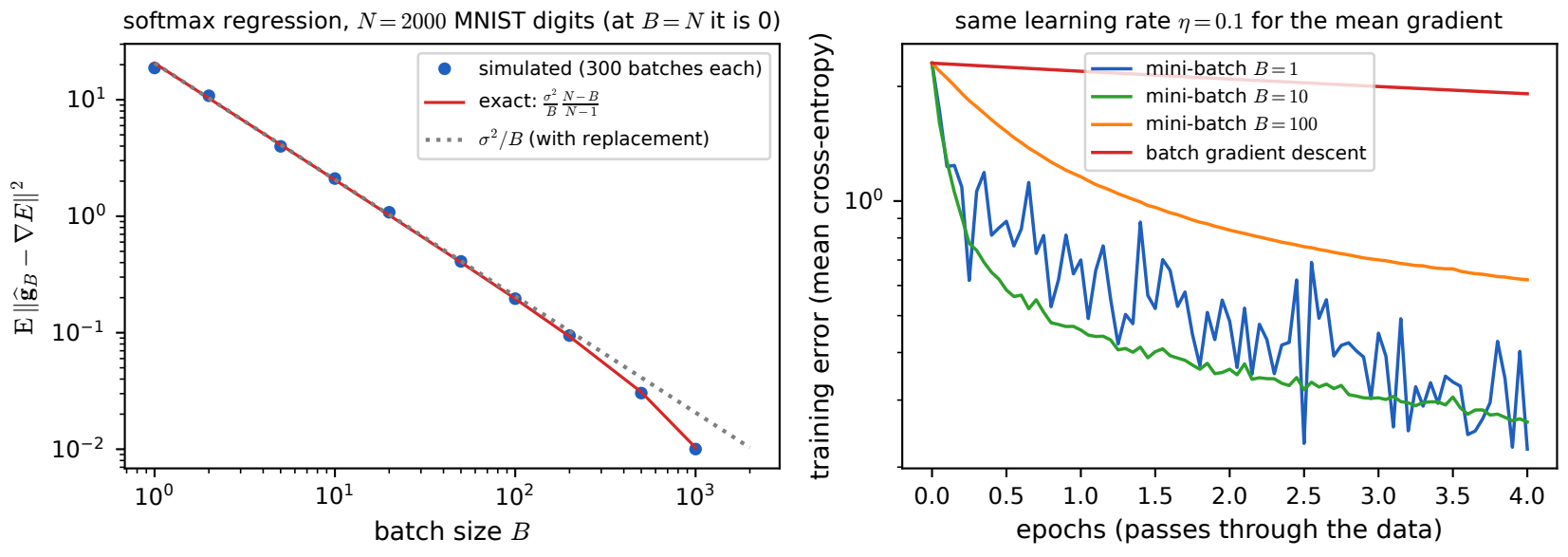

softmax regression, $N = 2000$ MNIST digits (at $B = N$ it is 0)
simulated (300 batches each)
exact: $\frac{\sigma^2}{B}\frac{N-B}{N-1}$
σ^2/B (with replacement)
$\mathrm{E}\,\|\hat{\mathbf{g}}_B - \nabla E\|^2$
batch size B
same learning rate $\eta = 0.1$ for the mean gradient
mini-batch $B = 1$
mini-batch $B = 10$
mini-batch $B = 100$
batch gradient descent
training error (mean cross-entropy)
epochs (passes through the data)