ADAM

Kingma and Ba (2015) — Adam

Published as a conference paper at ICLR 2015

ADAM: A METHOD FOR STOCHASTIC OPTIMIZATION

Diederik P. Kingma*
University of Amsterdam, OpenAI
dpkingma@openai.com

Jimmy Ba
University of Toronto
jimmy@psi.utoronto.ca

Under review as a conference paper at ICLR 2015

ABSTRACT

We introduce Adam, an algorithm for first-order gradient-based optimization of stochastic objective functions, based on adaptive estimates of lower-order moments. The method is straightforward to implement, is invariant to diagonal rescaling of the gradients, and is suitable for problems that are large in data and/or parameters. Adam is also appropriate for non-stationary objectives and problems with very noisy and/or sparse gradients. The hyper-parameters have intuitive interpretations and typically require little tuning. Some abstract theoretical analysis of the proposed method is also provided. Finally we provide a convergence proof for convex objectives, which is better than the known bounds.

1. INTRODUCTION

Stochastic gradient-based optimization is of core practical importance in many fields of science and engineering. Many problems in these fields can be cast as the optimization of some scalar parameter-ized objective function with respect to its parameters. If the objective function is differentiable with respect to its parameters, stochastic gradient descent is a relatively efficient approach (Section 2.4). Some of the most popular stochastic optimization algorithms are also the most simple: stochastic gradient descent (SGD) and momentum SGD. Recently, SGD has been used to train deep neural networks and convolutional neural networks (LeCun et al., 1998), and momentum SGD has been used to train deep auto-encoders (Hinton & Salakhutdinov, 2006). Other stochastic optimization methods that have been used in machine learning include Hessian-free optimization (Martens, 2010) and variants thereof (Martens & Sutskever, 2011), and methods that are based on the natural gradient (Amari, 1998) such as TONGA (Bordes et al., 2009) and Hessian-free optimization with a Kronecker-factored approximation (Martens & Grosse, 2015). Stochastic objectives arise in machine learning in the form of noisy estimates of the objective function. Many such noisy objectives are derived from the popular batch-wise approach, which replaces the full objective function with an unbiased estimate based on a single batch. Stochastic optimization methods that are robust to noise are clearly preferable in this setting, since they can be used with the most precise objectives, and stochastic gradient descent is often used with a single batch at each time step in practice. Stochastic objectives also arise in other ways, such as with dropout regularization (Hinton et al., 2012). We propose Adam, a method for efficient stochastic optimization that only requires first-order gradients with little memory requirement. The method computes adaptive estimates of lower-order moments, and is invariant to diagonal rescaling of the gradients. Section 2 describes the algorithm, and Section 3 describes its theoretical properties. Section 4 provides a convergence proof for convex objectives, and Section 5 provides a convergence analysis for non-convex objectives. Section 6 provides a brief review of related work, and Section 7 discusses some properties of Adam and its variants. Section 8