

CLAIMS

1. A computer-implemented method for training a model, comprising:
 - obtaining a training example for a model having model parameters stored on one or more computer readable storage mediums operably coupled to the hardware processor, the training example including an outcome and features to explain the outcome;
 - calculating a gradient with respect to the model parameters of the model using the training example;
 - computing at least two estimates of a moment of the gradient with at least two different time constants for the same type of the moment using the gradient; and
 - updating, using a hardware processor, the model parameters of the model using the at least two estimates of the moment with the at least two different time constants to reduce errors while calculating the at least two estimates of the moment of the gradient.
2. The method of claim 1, wherein each of the model parameters is updated with an amount determined individually by respective components of the at least two estimates of the moment in a conservative manner.
3. The method of claim 2, wherein a first model parameter of the model is updated by zero or a small amount in response to the at least two estimates of the moment being inconsistent in a component corresponding to the first model parameter.
4. The method of claim 3, wherein, in response to the at least two estimates of the moment being consistent in the component corresponding to the first model parameter, the first model parameter is updated according to a value generated by combining respective components of the at least two estimates of the moment corresponding to the first model parameter.
5. The method of any of claims 2 to 4, wherein a first model parameter of the model is updated according to a maximum or a mean of components of the at least two estimates of the moment corresponding to the first model parameter.
6. The method of any of the preceding claims, wherein the moment includes a first order moment and a second order moment as different types, wherein the first order moment represents average of the gradient and the second order moment scales individual learning rates for the model parameters of the model.
7. The method of any of claims 1 to 5, wherein the moment includes a first order moment and a second order moment as different types and a first model parameter of the model is updated in a manner depending on inconsistency between at least two estimates of the first order moment in a component corresponding to the first

model parameter and magnitude relationship between at least two estimates of the second order moment in the component.

8. The method of any of the preceding claims, wherein the time constants change exponential decay rates for moment estimation and the time constants include a first time constant and a second time constant that is larger or smaller than the first time constant.

9. The method of any of the preceding claims, wherein calculating the gradient, computing the at least two estimates of the moment, and updating the model parameters are iteratively performed in response to a new training example.

10. The method of any of the preceding claims, wherein the training example is provided in a streaming manner, wherein the model to be trained is updated each time a new training example arrives and the model is used to predict a value of the outcome based on input features.

11. The method of claim 10, wherein the input features include a plurality of elements representing past value fluctuations of the outcome observed over a predetermined period.

12. The method of claim 10, wherein the input features include a plurality of elements related to the outcome.

13. The method of any of the preceding claims, wherein the gradient is a stochastic gradient of an objective function at an iteration step, wherein the objective function evaluates a loss between the outcome in the training example and a prediction done by the model with current values of the model parameters from the features in the training example and the training example includes a single training example or a group of training examples.

14. A computer system for training a model by executing program instructions, the computer system comprising:

one or more computer readable storage mediums for storing the program instructions and a training example for a model having model parameters; and

processing circuitry in communication with the memory for executing the program instructions, wherein the processing circuitry is configured to:

obtain a training example for a model having model parameters from the one or more computer readable storage mediums, wherein the training example includes an outcome and features to explain the outcome;

calculate a gradient with respect to the model parameters of the model using the training example;

compute at least two estimates of a moment with at least two different time constants for the same type of the moment using the gradient; and

update the model parameters of the model using the at least two estimates of the moment with the at least two different time constants to reduce errors while calculating the at least two estimates of the moment of the gradient.

15. The computer system of claim 14, wherein the processing circuitry is configured to update each of the model parameters with an amount determined individually by respective components of the at least two estimates of the moment in a conservative manner.

16. The computer system of claim 15, wherein the processing circuitry is configured to update a first model parameter of the model by zero or a small amount in response to the at least two estimates of the moment being inconsistent in a component corresponding to the first model parameter and in response to the at least two estimates of the moment being consistent in the component corresponding to the first model parameter, wherein the first model parameter is updated according to a value generated by combining respective components of the at least two estimates of the moment corresponding to the first model parameter.

17. The computer system of claim 15, wherein the processing circuitry is configured to update a first model parameter of the model according to a maximum or a mean of components of the at least two estimates of the moment corresponding to the first model parameter.

18. The computer system of any of claims 14 to 17, wherein the moment includes a first order moment of the gradient and a second order moment of the gradient as different types, wherein the first order moment represents average of the gradient and the second order moment scales individual learning rates for the model parameters of the model.

19. The computer system of any of claims 14 to 18, wherein the moment includes a first order moment and a second order moment as different types and a first model parameter of the model is updated in a manner depending on inconsistency between at least two estimates of the first order moment in a component corresponding to the first model parameter and a magnitude relationship between at least two estimates of the second order moment in the component.

20. A computer program product for training a model, the computer program product comprising:
a computer readable storage medium readable by a processing circuit and storing instructions for execution by the processing circuit for performing a method according to any of claims 1 to 13.

21. A computer program stored on a computer readable medium and loadable into the internal memory of a digital computer, comprising software code portions, when said program is run on a computer, for performing the method of any of claims 1 to 13.