Term Paper: How and Why to Use Stochastic Gradient Descent?

Minjie Fan
Department of Statistics
University of California, Davis

Abstract

Stochastic gradient descent is a simple but efficient numerical optimization method. It has been widely used in solving large-scale machine learning problems. This paper first shows how to implement stochastic gradient descent, particularly for ridge regression and regularized logistic regression. Then the pros and cons of the method are demonstrated through two simulated datasets. The comparison of stochastic gradient descent with a state-of-the-art method L-BFGS is also done.

1 Introduction

Let us consider the following scenario: $(\boldsymbol{X}_i, Y_i), i = 1, \cdots, n$ are samples from a population $(\boldsymbol{X}, Y)$, where Y is a scalar, $\boldsymbol{X}$ can be either a scalar or a vector, and n is the sample size. Usually Y is called the response and $\boldsymbol{X}$ is called the predictor. A parametric predictive function $f_{\boldsymbol{\theta}}$ is used to predict Y based on $\boldsymbol{X}$, where $\boldsymbol{\theta}$ is the parameter vector. The predictive accuracy is measured by the loss function $l(\widehat{Y}, Y) = l(f_{\boldsymbol{\theta}}(\boldsymbol{X}), Y)$. The parameter vector $\boldsymbol{\theta}$ is estimated by minimizing the risk function

$$E(l(\widehat{Y}, Y)) = \int_{\Omega} l(f_{\boldsymbol{\theta}}(\boldsymbol{X}), Y) dP,$$

where P is the underlying probability measure. However, since the distribution of $(\boldsymbol{X}, Y)$ is unknown in practice, we usually approximate the risk by the empirical one

$$E_n(l(\widehat{Y}, Y)) = \frac{1}{n} \sum_{i=1}^{n} l(f_{\boldsymbol{\theta}}(\boldsymbol{X}_i), Y_i) = \frac{1}{n} \sum_{i=1}^{n} l_i(\boldsymbol{\theta}).$$

This form of objective function, i.e., $\sum_i l_i(\boldsymbol{\theta})$, is very common in supervised learning. Additionally, the minimizer of the objective function is called an M-estimator.

Traditional numerical optimization methods, such as those we have learned in class: gradient descent, Newton's method and BFGS, can be used to minimize the objective function. For example, at each iteration, gradient descent updates $\boldsymbol{\theta}$ as follows (assuming that the gradient of l_i exists)

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} - \alpha_t \sum_{i=1}^{n} \nabla l_i(\boldsymbol{\theta}^{(t)}),$$

where α_t is the step size (or called the learning rate). However, when the sample size is enormous, evaluating the gradient of the objective function, which is the sum of **all** the gradients of l_i, becomes time prohibitive. Given the limitation of the traditional methods, stochastic gradient descent (see Bottou (2012)) is proposed to speed up the computation through approximating the true gradient by the sum of a randomly selected subset of the gradients of l_i.

The contributions of this paper are: (i) showing how to implement stochastic gradient descent, particularly for ridge regression and regularized logistic regression; (ii) demonstrating the pros and

cons of stochastic gradient descent through comparing it with other traditional methods on simulated datasets.

The remainder of the paper is organized as follows. Section 2 introduces stochastic gradient descent. Its implementation on ridge regression and regularized logistic regression is shown in Section 3. We demonstrate stochastic gradient descent in Section 4 on simulated datasets. Section 5 concludes the paper.

2 Stochastic Gradient Descent

In its simplest form, stochastic gradient descent updates $\boldsymbol{\theta}$ as follows

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} - \alpha_t \nabla l_i(\boldsymbol{\theta}^{(t)}),$$

where the index i is randomly selected at each iteration. In practice, we usually randomly shuffle the samples and then go through them sequentially. Note that the method can be straightforwardly extended to using the gradient of more than one sample, i.e., $\sum_{j=1}^{b} \nabla l_{i+j}(\boldsymbol{\theta}^{(t)})$, which is called mini-batch gradient descent. There are some advantages of mini-batching, such as more stable convergence and suitable for faster matrix operations and parallelization.

When the step size decreases according to $\sum_t \alpha_t^2 < \infty$ and $\sum_t \alpha_t = \infty$, e.g., $\alpha_t = \mathcal{O}(1/t)$, and under certain mild conditions, stochastic gradient descent converges almost surely to a local minimum, and even a global minimum for a convex objective function. As we have learned in class, under certain regularity conditions, gradient descent and Newton's method can achieve linear convergence and quadratic convergence, respectively. This is equivalent to that to achieve an accuracy of ϵ for the objective function, gradient descent takes $\mathcal{O}(\log(1/\epsilon))$ iterations, and Newton's method takes even fewer. However, stochastic gradient descent takes $\mathcal{O}(1/\epsilon)$ iterations, which seems to be exponentially worse than gradient descent. Nonetheless, when n is sufficiently large, assuming that the time complexity of calculating the gradient of one sample is a constant C, the total time complexity of stochastic gradient descent is $\mathcal{O}(C/\epsilon)$, which is smaller than that of gradient descent, $\mathcal{O}(nC\log(1/\epsilon))$.

3 Implementation: Ridge Regression and Regularized Logistic Regression

In statistics and machine learning, regularization is usually used to solve an ill-posed problem or to prevent overfitting. The latter is originated from minimizing the empirical risk instead of the true one. Although stochastic gradient descent is designed for datasets with a large sample size, a probably large number of features make it necessary to incorporate regularization into the empirical risk to reduce the generalization error. In the sequel, we shall present the implementation of stochastic gradient descent on ridge regression and regularized logistic regression with ℓ_2-norm. Note that the method is not restricted to these two problems. Other problems, such as Lasso (Shalev-Shwartz and Tewari, 2011), support vector machines (Menon, 2009) and neural networks (Bottou, 1991) can be solved by stochastic gradient descent or its variants.

3.1 Ridge Regression

Ridge regression is formulated as

$$(\hat{\beta}_0, \hat{\boldsymbol{\beta}}) = \arg\min_{\beta_0, \boldsymbol{\beta}} \frac{1}{2n} \sum_{i=1}^{n} (y_i - \beta_0 - \boldsymbol{x}_i^T \boldsymbol{\beta})^2 + \frac{\lambda}{2} \|\boldsymbol{\beta}\|_2^2,$$

where $\boldsymbol{x}_i^T$ is the i-th row vector of the $n \times p$ (p is the number of features) design matrix $\boldsymbol{X}$, $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)^T$, and λ is the tuning parameter that controls the regularization. It is clear that for all $\lambda > 0$, the objective function is strictly convex, which implies that its minimizer exists and is unique. We write the objective function as

$$(\hat{\beta}_0, \hat{\boldsymbol{\beta}}) = \arg\min_{\beta_0, \boldsymbol{\beta}} \frac{1}{n} \sum_{i=1}^{n} l_i(\boldsymbol{\beta}),$$

where $l_i(\boldsymbol{\beta}) = \frac{1}{2}(y_i - \beta_0 - \boldsymbol{x}_i^T\boldsymbol{\beta})^2 + \frac{\lambda}{2}\|\boldsymbol{\beta}\|_2^2$. The gradient of l_i is

$$\nabla l_i(\boldsymbol{\beta}) = (-y_i + \beta_0 + \boldsymbol{x}_i^T\boldsymbol{\beta})(1, \boldsymbol{x}_i^T)^T + \lambda(0, \boldsymbol{\beta}^T)^T. \tag{1}$$

The minimizer of the objective function can be obtained by setting $\sum_i \nabla l_i(\boldsymbol{\beta})$ as 0. W.l.o.g., assuming $\beta_0 = 0$, the minimizer has the closed form

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X} + \lambda \boldsymbol{I}_n)^{-1}\boldsymbol{X}^T\boldsymbol{y},$$

where $\boldsymbol{y} = (y_1, \cdots, y_n)^T$ is the vector of responses. However, the time complexity of inverting (or more exactly, calculating its Cholesky decomposition) a $p \times p$ matrix is $\mathcal{O}(p^3)$, which becomes time prohibitive when p is enormous. In comparison, the time complexity of stochastic gradient descent is $\mathcal{O}(p/\epsilon)$, which will be smaller than $\mathcal{O}(p^3)$ if the number of iterations is less than $\mathcal{O}(p^2)$. Note that here we do not consider the case when $\boldsymbol{X}$ is sparse.

3.2 Regularized Logistic Regression

Regularized logistic regression is formulated as

$$\begin{aligned}(\hat{\beta}_0, \hat{\boldsymbol{\beta}}) &= \arg\min_{\beta_0, \boldsymbol{\beta}} -\frac{1}{n}\log(\text{likelihood function}) + \frac{\lambda}{2}\|\boldsymbol{\beta}\|_2^2 \\ &= \arg\min_{\beta_0, \boldsymbol{\beta}} \frac{1}{n}\sum_{i=1}^{n}\left[\log\left\{1 + \exp(\beta_0 + \boldsymbol{x}_i^T\boldsymbol{\beta})\right\} - y_i(\beta_0 + \boldsymbol{x}_i^T\boldsymbol{\beta})\right] + \frac{\lambda}{2}\|\boldsymbol{\beta}\|_2^2.\end{aligned}$$

For all $\lambda > 0$, given that its Hessian matrix is positive definite, the objective function is strictly convex, which implies that its minimizer exists and is unique. We write the objective function as

$$(\hat{\beta}_0, \hat{\boldsymbol{\beta}}) = \arg\min_{\beta_0, \boldsymbol{\beta}} \frac{1}{n}\sum_{i=1}^{n} l_i(\boldsymbol{\beta}),$$

where $l_i(\boldsymbol{\beta}) = \log\left[1 + \exp(\beta_0 + \boldsymbol{x}_i^T\boldsymbol{\beta})\right] - y_i(\beta_0 + \boldsymbol{x}_i^T\boldsymbol{\beta}) + \frac{\lambda}{2}\|\boldsymbol{\beta}\|_2^2$. The gradient of l_i is

$$\nabla l_i(\boldsymbol{\beta}) = [-y_i + S(\beta_0 + \boldsymbol{x}_i^T\boldsymbol{\beta})](1, \boldsymbol{x}_i^T)^T + \lambda(0, \boldsymbol{\beta}^T)^T, \tag{2}$$

where $S(\cdot)$ denotes the sigmoid function, i.e.,

$$S(x) = \frac{\exp(x)}{1 + \exp(x)}.$$

An interesting observation is that the gradients (1) and (2) share the same form except that the sigmoid function in (2) is replaced by the identity function in (1).

3.3 Implementation

For both ridge regression and regularized logistic regression, the tuning parameter λ can be determined by cross-validation. In the sequel, we shall assume that λ is known and fixed since our focus is on minimizing the objective function.

As one of the disadvantages of stochastic gradient descent, choosing a proper step size is challenging. We simply use the form suggested by Bottou (2012)

$$\alpha_t = \frac{\alpha_0}{1 + \alpha_0\lambda t},$$

where t is the iteration number starting from 0 and α_0 is the initial step size, which can be tuned using a small subset of the dataset. See Schaul *et al.* (2013) for an adaptive method that does not need any manual tuning but compares favorably with an "ideal SGD".

By default, the initial value of $(\beta_0, \boldsymbol{\beta})$ is a zero vector. For a warm start, one may specify it as the estimate obtained by other methods using a small subset of the dataset. The total number of iterations is set as $1e6$ according to Pedregosa *et al.* (2011), which empirically ensures the convergence of stochastic gradient descent. All my codes are written in *Julia*. It is worth pointing out that there is only one *Julia* package available for stochastic gradient descent, which is called *SGDOptim*.

4 Simulation Study

This section applies stochastic gradient descent to two simulated datasets: one for ridge regression and the other for regularized logistic regression.

4.1 Simulated Dataset 1: Ridge Regression

We simulate data from a linear model with i.i.d. Gaussian random errors. The coefficient vector $(\beta_0, \boldsymbol{\beta}) = (3, -4, 5)$. The standard deviation of the random errors is 0.1. All the entries of the design matrix $\boldsymbol{X}$ are simulated from $\mathcal{N}(0, 1)$ independently. Since the aim of this simulation study is to check how stochastic gradient descent converges to the global minimum, we only specify n as 100. Given that the number of features is only 2, the tuning parameter λ is fixed at $1e{-}4$. By trial and error, we find $\alpha_0 = 0.3$ to be suitable. From Figure 1, we can see that the objective function moves close to the minimum very fast, and then oscillates around it. As the iteration number increases, the pronounced oscillation at the very beginning is gradually damped to be negligible. This confirms the aforementioned almost sure convergence of stochastic gradient descent. It is notable that the convergence to the minimum takes a very long time. In practice, one may use fewer iterations when the running time is the bottleneck.

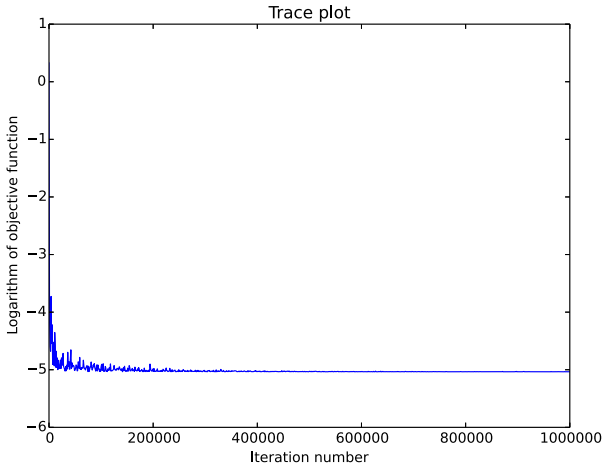

Figure 1: Trace plot of the logarithm of the objective function (only every 1000 iterations).

4.2 Simulated Dataset 2: Regularized Logistic Regression

We turn to compare the computational speed of stochastic gradient descent and L-BFGS on a simulated dataset from a logistic regression model. L-BFGS is a limited-memory version of BFGS. It achieves super-linear convergence with the cost per iteration smaller than Newton's method. Due to these advantages, it has been a popular algorithm for parameter estimation in machine learning. We use the *Julia* package *Optim* to run BFGS.

The experiment is done under a high-dimensional setting, where $n = 1e6$ and $p = 100$. All the entries of the coefficient vector and the design matrix are simulated from $\mathcal{N}(0, 1)$ independently. The tuning parameter λ is fixed at $1e - 4$. By trial and error, we find $\alpha_0 = 0.005$ to be suitable.

In terms of the computational speed, stochastic gradient descent takes 9.45 seconds, while L-BFGS takes 151.49 seconds, which is much slower than the former. Figure 2 contains the comparison among the true coefficients and the estimated coefficients by stochastic gradient descent and L-BFGS. The RMSE of these two estimated coefficients are almost the same, which are 0.04926 and 0.04924, respectively. Thus, stochastic gradient descent and L-BFGS achieve the equivalent accuracy but the former performs almost 20 times faster than the latter.

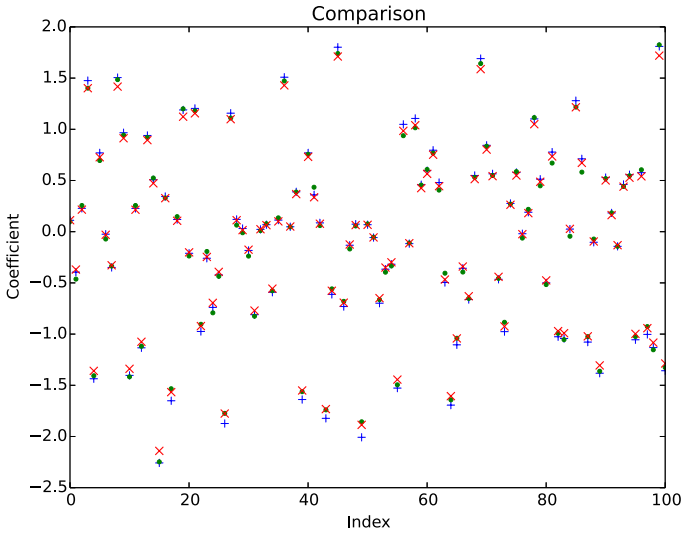

Figure 2: Comparison among the true coefficients (+) and the estimated coefficients by stochastic gradient descent (.) and L-BFGS (x).

5 Conclusion

In this paper, we have shown how to implement stochastic gradient descent. We have also demonstrated the superiority of stochastic gradient descent to other traditional methods for large-scale machine learning problems. It is efficient and easy to implement, but tuning the step size parameter needs some efforts.

References

Bottou, L. (1991) Stochastic gradient learning in neural networks. *Proceedings of Neuro-Nımes*, **91**.

Bottou, L. (2012) Stochastic gradient descent tricks. In *Neural Networks: Tricks of the Trade*, 421–436. Springer.

Menon, A. K. (2009) Large-scale support vector machines: algorithms and theory. *Research Exam, University of California, San Diego*, 1–17.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. and Duchesnay, E. (2011) Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, **12**, 2825–2830.

Schaul, T., Zhang, S. and Lecun, Y. (2013) No more pesky learning rates. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, 343–351.

Shalev-Shwartz, S. and Tewari, A. (2011) Stochastic methods for ℓ_1-regularized loss minimization. *The Journal of Machine Learning Research*, **12**, 1865–1892.