Continuous-Time Limit of Stochastic Gradient Descent Revisited

Stephan Mandt
Columbia University
sm3976@columbia.edu

Matthew D. Hoffman
Adobe Research
mathoffm@adobe.com

David M. Blei
Columbia University
david.blei@columbia.edu

Abstract

Stochastic Gradient Descent (SGD) is an important algorithm in machine learning. With constant learning rates, it is a stochastic process that reaches a stationary distribution. We revisit an analysis of SGD in terms of stochastic differential equations in the limit of small constant gradient steps. This limit, which we feel is not appreciated in the machine learning community, allows us to approximate SGD in terms of a multivariate Ornstein-Uhlenbeck process, and hence to compute stationary distributions in closed form. This formalism has interesting new implications for machine learning. We consider the case where the objective has the interpretation of a log-posterior. Traditional theory suggests choosing the learning rate such that the stationary distribution approximates a point mass at the optimum, but this can lead to wasted effort and overfitting. When the goal is instead to approximate the posterior as well as possible, we can derive criteria for optimal minibatch sizes, learning rates, and preconditioning matrices.

1 Introduction

Stochastic gradient descent (SGD) is one of the most widely used optimization algorithms in machine learning [1]. Classical SGD involves a decreasing learning rate, which guarantees convergence to an optimum [2]. Recently, however, there has been growing interest in constant learning rates [3, 4], even though the resulting algorithm does not converge in the classical sense. Rather, the iterates of SGD with constant learning rates converge *in distribution* to a stationary density over parameter space. This stationary distribution has its maximum at the optimum of the objective function, but has a non-zero variance. If we cannot reliably estimate our parameters beyond the variance of the stationary distribution, then a sample from this distribution may be adequate [5].

The properties of SGD have traditionally been analyzed using quasi-martingales [6, 7]. In this paper, we review an alternative perspective on constant-rate SGD and an alternative way to analyze its properties [8]. We first derive the multivariate Ornstein-Uhlenbeck process [9], a continuous-time stochastic process that approximates SGD in the vicinity of a local optimum. This allows us to explicitly calculate stationary distributions. These distributions depend on various parameters of the optimization problem—the learning rate, the batch size, and optionally preconditioning matrices. We reinterpret the stationary distribution as an approximation of a posterior distribution [10], where the parameters play the role of variational parameters. We can tune these parameters to optimally fit the sampling distribution of the iterates to the posterior. This yields new criteria for setting the parameters of SGD. As a result, we derive preconditioners that relate to AdaGrad [11] and stochastic gradient Fisher scoring [12], as well as early stopping criteria. We also show that when using the formalism of stochastic differential equations, one can give a fast derivation of a result by Polyak [13] that iterate averaging is unaffected by preconditioning the stochastic gradient.

The approach can be also applied to variations of the SGD algorithm, giving new heuristics for analyzing asymptotic behavior.

Related Work. There is abundant work on the convergence behavior of stochastic gradient descent [1, 7]. Many papers discuss constant step-size SGD. [4, 14] discuss convergence rate of averaged gradients with constant step size, while [3] analyze sampling distributions using quasi-martingale techniques. In our paper, we also further investigate results by [13]. Ways of analyzing SGD in terms of stochastic differential equations and discrete Ornstein-Uhlenbeck processes can be also found in [8, 15] and are here reviewed and extended. Our theoretical analysis is also related to Bayesian statistics, in particular Stochastic Gradient Langevin Dynamics (SGLD) by [16, 17], where posterior sampling is achieved by adding artificial noise to the stochastic gradient. See also [12, 18, 19] for similar stochastic gradient-based MCMC algorithms. These works focus on (asymptotically) exact sampling from a Bayesian posterior by averaging away SGD sampling noise. In contrast, our goal is to explore the properties of the sampling noise and its influence on the stationary distribution of SGD. [10] also interpret SGD as a non-parametric variational inference scheme, but with different goals and in a very different formalism.

2 The Continuous-Time Approximation Revisited

Problem setup and assumptions. The asymptotic properties of stochastic gradient descent can be analyzed as a continuous-time stochastic process [8]. This formalism is not appreciated in the machine learning community and shall here be revisited and applied for constant learning rates. We consider objective functions of the form $\mathcal{L}(\theta) = \sum_{n=1}^{N} \ell_n(\theta)$. Let $\mathcal{S}(t)$ be a set of S random indices drawn uniformly at random from the set $\{1, \cdots, N\}$. $\mathcal{S}$ indexes the functions $\ell_n(\theta)$ that are subsampled in each iteration. S is the minibatch size. We can form a stochastic estimate of the objective and a stochastic gradient,

$$\hat{\mathcal{L}}(\theta) = \frac{N}{S} \sum_{n \in \mathcal{S}} \ell_n(\theta), \quad \hat{g}_S(\theta) = \nabla_\theta \hat{\mathcal{L}}(\theta). \tag{1}$$

In expectation the stochastic gradient is the full gradient, $g(\theta) = \mathbb{E}[\hat{g}_S(\theta)]$. This stochastic gradient is then used for the SGD update, involving a constant and small learning rate ϵ:

$$\theta(t+1) \quad = \quad \theta(t) - \epsilon\, \hat{g}_S(\theta(t)). \tag{2}$$

Equations 1 and 2 are the real (discrete time) process that SGD follows. We will approximate it with a continuous time process and then analyze it. We first state some **assumptions**.

- A1 Assume that the gradient noise $\hat{g}_S(\theta) - g(\theta)$ is Gaussian distributed.
- A2 Assume that the iterates $\theta(t)$ are constrained to a small enough region in parameter space that the sampling noise covariance of the stochastic gradients is constant.
- A3 Assume that the step size is small enough that we can approximate the discrete-time Markov chain defined by the SGD algorithm with a continuous-time Markov process.

Below, we will use these assumptions to derive a class of stochastic processes that approximate stochastic gradient descent. We will also discuss the plausibility of these assumptions.

Mapping to stochastic differential equations. Starting with Eq. 2, we first approximate the sampling noise. Assumption [A1] is based on idea that the stochastic gradient is an average of S independent random contributions. If S is sufficiently large, the central limit theorem should apply, making $\hat{g}_S$ Gaussian with variance $\propto 1/S$. Hence, we can write the stochastic gradient as

$$\hat{g}_S(\theta) \quad \approx \quad g(\theta) + \hat{\xi}_S(\theta), \quad \hat{\xi}_S(\theta) \sim \mathcal{N}(0, C(\theta)/S). \tag{3}$$

This involves the stochastic gradient noise covariance $C(\theta)/S \equiv \mathbb{E}\left[(\hat{g}_S(\theta) - g(\theta))(\hat{g}_S(\theta) - g(\theta))^\top\right]$. The covariance matrix C, that we assumed to be independent of θ in the region of interest [A2], is positive definite and symmetric, and has a square root decomposition $C = BB^\top$. We introduced a rescaled noise covariance matrix $B_{\epsilon/S} = \sqrt{\epsilon/S}\, B$. We can now write the stochastic process as

$$\theta(t+1) - \theta(t) \;=\; -\epsilon\, g(\theta(t)) + \sqrt{\epsilon}\, B_{\epsilon/S}\, W(t), \quad W(t) \;\sim\; \mathcal{N}(0, \mathbf{I}). \tag{4}$$

We have re-written Eq. 2 in such a way that it assumes the form of a time discretization of a continuous-time stochastic differential equation. Based on assumption [A3] we can derive

$$d\theta(t) \quad = \quad -\nabla_\theta \mathcal{L}(\theta) dt + B_{\epsilon/S}\, dW(t) \tag{5}$$

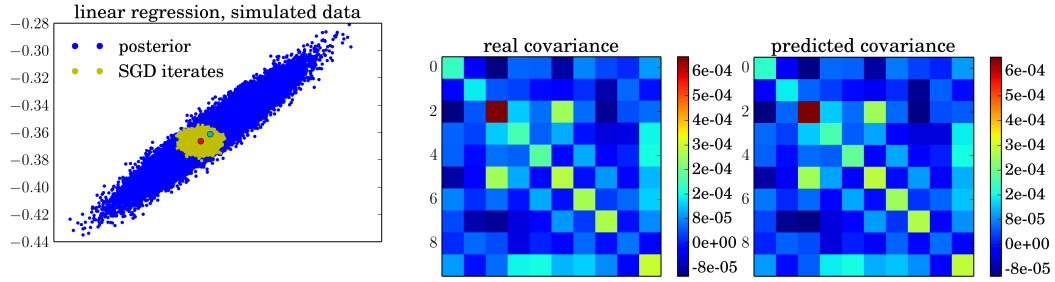

Figure 1: **Left:** Posterior distribution $f(\theta) \propto \exp\{-\mathcal{L}(\theta)\}$ (blue) and sampling distribution $q(\theta)$ of SGD (yellow) for a linear regression model as discussed in Section 3.3. The plot shows that the sampling distribution is approximately Gaussian and approximately centered around the maximum of the posterior. **Right:** Same experiments in 10 dimensions, where we checked the relation between the covariance of the sampling distribution and its prediction in terms of the Ornstein-Uhlenbeck process (Eq. 7). The good agreement between the middle and right plots supports our modeling assumptions.

as an approximation to SGD. This can be considered a generalized Langevin equation, and therefore we renamed $g(\theta)$ back into $\nabla_\theta \mathcal{L}(\theta)$. But note the usual Langevin equation assumes isotropic noise.

Above, we have performed the conventional substitution rules when discretizing a continuous-time stochastic process $\theta(t+1)-\theta(t) \to d\theta(t)$, $\epsilon \to dt$ and $\sqrt{\epsilon}\, W \to dW$, see e.g. [20]. The last substitution is required to respect the Wiener noise relation $\mathbb{E}[dW(t)dW(t')^\top] = \mathbf{I}\,\delta(t-t')dt$, which for finite ϵ corresponds to $\mathbb{E}[\sqrt{\epsilon}W(t)\sqrt{\epsilon}W(t')^\top] = \mathbf{I}\,\delta(t-t')\epsilon$. This equation is still not tractable because the noise is non-isotropic. We need a **further assumption**:

A4 Assume that the stationary distribution of the iterates is constrained to a region within which the objective is well approximated by a quadratic function.

Multivariate Ornstein-Uhlenbeck process. Assumptions [A1-A4] induce a multivariate *Ornstein-Uhlenbeck* process [9], which is our core modeling framework. Without loss of generality, assume the minimum is at $\theta = 0$, so $\mathcal{L}(\theta) = \frac{1}{2}\theta^\top A\theta$. This yields

$$d\theta(t) \quad = \quad -A\theta(t)dt + B_{\epsilon/S}\,dW(t) \tag{6}$$

This process has an analytic solution in terms of the stochastic integral [20], $\theta(t) = \exp(-At)\theta(0) + \int_0^t \exp[-A(t-t')]B_{\epsilon/S}\,dW(t')$. It has a Gaussian stationary distribution. To calculate the Gaussian covariance $\Sigma = \mathbb{E}[\theta(t)\theta(t)]$ we use a result from [20] (also contained in the appendix of [15]):

$$q(\theta) \quad \propto \quad \exp\left\{-\tfrac{1}{2}\theta^\top \Sigma^{-1}\theta\right\}, \quad \Sigma A^\top + A\Sigma = \tfrac{\epsilon}{S}BB^\top. \tag{7}$$

This distribution depends on the parameters of SGD and on properties of the objective and noise covariance. We can thus use it to analyze the behavior of SGD under our assumptions.

3 Implications for Machine Learning

3.1 Stochastic gradient as approximate inference

Let us assume that our optimization problem arises from fitting a joint distribution $p(\theta, x)$, involving data x and parameters θ. Hence $\mathcal{L}(\theta) \equiv -\log p(\theta, x)$. This is the case for many models, including neural networks. According to assumption [A4], the corresponding posterior $f(\theta) \equiv p(\theta|x)$ is

$$f(\theta) \quad \propto \quad \exp\left\{-\tfrac{1}{2}\theta^\top A\theta\right\}. \tag{8}$$

The classical goal of SGD, namely to minimize the objective, would lead to a point estimate $\theta^* = 0$ of the posterior. To avoid overfitting, an alternative goal could be to instead tune the parameters of SGD such that the stationary distribution approximates the posterior:

$$\{\epsilon^*, S^*, H^*\} \quad = \quad \min_{\epsilon,S,H} KL(q(\theta)\|f(\theta)). \tag{9}$$

This involves the learning rate ϵ, minibatch size S, and preconditioning matrix H. We first consider the basic SGD update (without preconditioning). Since $f(\theta)$ and $q(\theta)$ both are Gaussian with means 0, we can compute their KL divergence analytically:

$$KL(q\|f) \;=\; \mathbb{E}_{q(\theta)}[\log f(\theta)] - \mathbb{E}_{q(\theta)}[\log q(\theta)] \;=\; \tfrac{1}{2}\left(\mathrm{Tr}(A\Sigma) - \log|A| - \log|\Sigma| - d\right).$$

Above, $|\cdot|$ denotes the determinant. We can use Eq. 7 to simplify this expression which leads to $KL(q\|f) = \frac{\epsilon}{2S}\mathrm{Tr}(BB^\top) - \log(\epsilon/S) + const.$ Minimizing KL divergence over ϵ/S, we find $\epsilon^* = 2S/\mathrm{Tr}(BB^\top)$ for the optimal learning rate, which tells us that the constant learning rate should be inversely proportional to the sampling noise. This results gets more interesting if we precondition the stochastic gradient with a diagonal matrix H as a substitute for a scalar learning rate. Up to constants, the modified KL divergence is $KL(q\|f) = \frac{\epsilon}{2S}\mathrm{Tr}(BB^\top H) + \frac{1}{2}\log\left(\frac{\epsilon}{S}|H\Sigma^{-1}H|\right) = \frac{\epsilon}{2S}\mathrm{Tr}(BB^\top H) + \mathrm{Tr}\log(H) + \frac{1}{2}\log\frac{\epsilon}{S} - \log|\Sigma|$. If we constrain ourselves to diagonal preconditioners, we find

$$H_k^* \quad\propto\quad \frac{1}{2(BB^\top)_{kk}} \tag{10}$$

to be optimal. Hence, the optimal diagonal preconditioner is the inverse of the diagonal part of the noise matrix. This result relates to AdaGrad [11], which also adjusts the preconditioner to diagonal entries of the noise covariance, but note there are differences, e.g. our result does not suggest taking square roots. It also relates closely to stochastic gradient Fisher scoring [12], but does not require dense preconditioners, and applies to misspecified models.

3.2 Optimality of iterate averaging

One of the most important strategies for improving convergence rates in SGD is iterate averaging [13]. If our goal is to estimate the minimum of $\mathcal{L}$ using a sequence of stochastic gradients, then a natural approach is to compute an average of the iterates. After T steps of size ϵ and assuming [A3], this average is

$$\hat{\mu} \equiv \tfrac{1}{T}\textstyle\sum_{t=1}^{T}\theta(t\epsilon) \approx \tfrac{1}{T\epsilon}\int_0^{T\epsilon}\theta(t)dt \equiv \hat{\mu}' \tag{11}$$

The average $\hat{\mu}$ and its approximation $\hat{\mu}'$ are random variables whose expected value is the minimum of the objective. The accuracy of this estimator $\hat{\mu}'$ after a fixed amount of iterations T is determined by its covariance matrix. We can calculate this covariance using stochastic calculus and find

$$\mathbb{E}\left[\hat{\mu}'\hat{\mu}'^\top\right] \approx \tfrac{1}{TS}A^{-1}BB^\top(A^{-1})^\top. \tag{12}$$

This covariance depends only on the number of iterations T times the minibatch size S, not on the step size ϵ. Since TS is the total number of examples that are processed, this means that this iterate averaging scheme's efficiency does not depend on either the minibatch size or the step size, proven first in [13]. We can make a slightly stronger statement. If we precondition the stochastic gradients with a positive-definite matrix H, it turns out that the covariance of the estimator remains unchanged. The resulting process $d\theta = -HA\theta(t)dt + HB_{\epsilon/S}dW(t)$ is again of Ornstein-Uhlenbeck type. When replacing $A' \equiv HA$ and $B' \equiv HB$ and using Eq. 11, we see that H cancels out and the resulting distribution is the same. The stochastic formalism offers an explanation. Preconditioning can make the stationary distribution of SGD more isotropic. However, it also enhances the mixing times in the directions that are problematic. These effects exactly cancel each other.

3.3 Experiments

We generated $N = 10000$ data points from a linear regression model in $D = 10$ dimensions. We then ran SGD for $50,000$ iterations on a linear regression model with a quadratic penalty of $\frac{1}{20000}\|\theta\|^2$, a constant learning rate of $\epsilon = 0.01$, and a minibatch size of $S = 500$.

Fig. 1 (left) shows a two-dimensional projection of the sampling distribution (yellow) and samples from the model posterior (blue). The plot reveals that the sampling distribution is indeed approximately Gaussian and centered around the maximum of the posterior. To check our theoretical assumptions, we compared the covariance of the sampling distribution against its predicted value based on the Ornstein-Uhlenbeck process (Eq. 7). This is also shown in Fig. 1 (right). The two plots that compare the entries of the covariance matrices on a color scale show that the prediction is satisfied to a very good extent. This suggests that the assumptions of modeling SGD as an Ornstein-Uhlenbeck process are indeed satisfied for small learning rates. We saw similar behavior for logistic regression and linear regression on real-world data (not shown here).

References

[1] Léon Bottou. Online learning and stochastic approximations. *On-line learning in neural networks*, 17(9):25, 1998.

[2] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

[3] Alexandre Défossez and Francis Bach. Averaged least-mean-squares: Bias-variance trade-offs and optimal sampling distributions. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, pages 205–213, 2015.

[4] Francis Bach and Eric Moulines. Non-strongly-convex smooth stochastic approximation with convergence rate o (1/n). In *Advances in Neural Information Processing Systems*, pages 773–781, 2013.

[5] Olivier Bousquet and Léon Bottou. The tradeoffs of large scale learning. In *Advances in neural information processing systems*, pages 161–168, 2008.

[6] Léon Bottou. Stochastic learning. In *Advanced lectures on machine learning*, pages 146–168. Springer, 2004.

[7] Tong Zhang. Solving large scale linear prediction problems using stochastic gradient descent algorithms. In *Proceedings of the twenty-first international conference on Machine learning*, page 116. ACM, 2004.

[8] Harold J Kushner and George Yin. *Stochastic approximation and recursive algorithms and applications*, volume 35. Springer Science & Business Media, 2003.

[9] George E Uhlenbeck and Leonard S Ornstein. On the theory of the brownian motion. *Physical review*, 36(5):823, 1930.

[10] Dougal Maclaurin, David Duvenaud, and Ryan P Adams. Early stopping is nonparametric variational inference. *arXiv preprint arXiv:1504.01344*, 2015.

[11] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *The Journal of Machine Learning Research*, 12:2121–2159, 2011.

[12] Sungjin Ahn, Anoop Korattikara, and Max Welling. Bayesian posterior sampling via stochastic gradient fisher scoring. *arXiv preprint arXiv:1206.6380*, 2012.

[13] Boris T Polyak and Anatoli B Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, 1992.

[14] Nicolas Flammarion and Francis Bach. From averaging to acceleration, there is only a step-size. *arXiv preprint arXiv:1504.01577*, 2015.

[15] Lennart Ljung, Georg Ch Pflug, and Harro Walk. *Stochastic approximation and optimization of random systems*, volume 17. Birkhäuser, 2012.

[16] Max Welling and Yee W Teh. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 681–688, 2011.

[17] Issei Sato and Hiroshi Nakagawa. Approximation analysis of stochastic gradient langevin dynamics by using fokker-planck equation and ito process. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 982–990, 2014.

[18] Tianqi Chen, Emily B Fox, and Carlos Guestrin. Stochastic gradient hamiltonian monte carlo. *arXiv preprint arXiv:1402.4102*, 2014.

[19] Yi-An Ma, Tianqi Chen, and Emily B Fox. A complete recipe for stochastic gradient mcmc. *arXiv preprint arXiv:1506.04696*, 2015.

[20] Crispin W Gardiner et al. *Handbook of stochastic methods*, volume 4. Springer Berlin, 1985.