Stochastic Gradient Descent as Approximate Bayesian Inference

Skorik Sergey

MIPT, 2022

October 23, 2022

Problem setup

Consider loss functions of the following form:

$$\mathcal{L}(\theta) = \frac{1}{N}\sum_{n=1}^{N}\ell_n(\theta), \quad g(\theta) \equiv \nabla_\theta L(\theta), \;\; \ell_n(\theta) \equiv \ell(\theta, \boldsymbol{x}_n) \Rightarrow \mathcal{L}(\theta) \equiv \mathcal{L}(\theta, \boldsymbol{x}) \tag{1}$$

Let $\mathcal{S}$ be a set of S random indices drawn uniformly at random from the set $\{1, ..., N\}$. we call $\mathcal{S}$ a "minibatch" of size S.
Then, let's define

$$\hat{\mathcal{L}}_S(\theta) = \frac{1}{S}\sum_{n\in\mathcal{S}}\ell_n(\theta), \quad \hat{g}_S(\theta) \equiv \nabla_\theta\hat{\mathcal{L}}_S(\theta), \;\; g(\theta) = \mathbb{E}[\hat{g}_S(\theta)] \tag{2}$$

We use this stochastic gradient in the SGD update

$$\theta(t+1) = \theta(t) - \varepsilon\hat{g}_S(\theta(t)), \quad \varepsilon = const \tag{3}$$

Eqs. 2 and 3 define the discrete-time process that SGD simulates from. We will approximate it with a continuous-time process that is easier to analyze.

SGD as an Ornstein-Uhlenbeck Process

Definitions

Itô stochastic differential equation:

$$\begin{cases} dX(t) = f\left(t, X(t)\right) dt + g\left(t, X(t)\right) dW(t) \\ X(0) = X_0 \end{cases} \tag{4}$$

Definitions

The Ornstein–Uhlenbeck process $X(t)$ is defined by the following stochastic differential equation:

$$dX(t) = \theta X(t) dt + \sigma dW(t) \tag{5}$$

where $\theta > 0$, $\sigma > 0$ are parameters and $W(t)$ denotes the Wiener process.

We now show how to approximate the discrete-time Eq. 3 with a continuous-time Ornstein-Uhlenbeck process. To justify the approximation, we make four assumptions.

SGD as an Ornstein-Uhlenbeck Process

Assumptions

Assumption 1: *Observe that the stochastic gradient is a sum of S independent, uniformly sampled contributions. Then, we apply the central limit theorem*

$$\hat{g}_S(\theta(t)) = \sum_{i=1}^{S} \hat{g}_i(\theta(t)), \quad \hat{g}_i(\theta(t)) - i.i.d, \ \mathbb{E}[\hat{g}_i(\theta)] = \frac{1}{S} g(\theta), \ i = \overline{1, n}$$

$$\Delta g(\theta) = \frac{\hat{g}_S(\theta(t)) - S\frac{1}{S} g(\theta)}{\sqrt{S}} \xrightarrow{\mathbb{P}} \mathcal{N}\left(0, \frac{1}{S} C(\theta)\right)$$

Hence

$$\hat{g}_S(\theta(t)) \approx g(\theta) + \frac{1}{\sqrt{S}} \Delta g(\theta), \quad \Delta g(\theta) \sim \mathcal{N}(0, C(\theta)) \tag{6}$$

SGD as an Ornstein-Uhlenbeck Process

Assumptions

Assumption 2: *We assume that the covariance matrix $C(\theta)$ is approximately constant with respect to θ. As a symmetric positive-semidefinite matrix, this constant matrix C factorizes as*

$$C(\theta) \approx C = BB^\top \tag{7}$$

We now define $\Delta\theta(t) = \theta(t+1) - \theta(t)$ and combine Eqs. 3, 6 and 7 to rewrite process as

$$\Delta\theta(t) = -\varepsilon g(\theta(t)) + \frac{\varepsilon}{\sqrt{S}} B\Delta W, \quad \Delta W \sim \mathcal{N}(0, \boldsymbol{I}) \tag{8}$$

Assumption 3: *We assume that we can approximate the finite-difference equation (8) by the stochastic differential equation*

$$d\theta(t) = -\varepsilon g(\theta(t))dt + \frac{\varepsilon}{\sqrt{S}} BdW(t) \tag{9}$$

SGD as an Ornstein-Uhlenbeck Process

Assumptions

Assumption 4: *We assume that the stationary distribution of the iterates is constrained to a region where the loss is well approximated by a quadratic function.*

$$\mathcal{L}(\theta) = \frac{1}{2}\theta^{\top} A\theta \tag{10}$$

(Without loss of generality, we assume that a minimum of the loss is at $\theta = 0$*.) We also assume that* A *is positive definite.*

The four assumptions above result in a specific kind of stochastic process, the multivariate Ornstein-Uhlenbeck process.

$$d\theta(t) = -\varepsilon A\theta(t)dt + \frac{\varepsilon}{\sqrt{S}} BdW(t) \tag{11}$$

SGD as an Ornstein-Uhlenbeck Process

Discussion

This connection helps us analyze properties of SGD because the Ornstein-Uhlenbeck process has an analytic stationary distribution $q(\theta)$ that is Gaussian.

$$q(\theta) \propto \exp\left(-\frac{1}{2}\theta^\top \Sigma^{-1}\theta\right) \tag{12}$$

The covariance Σ satisfies

$$\Sigma A + A\Sigma = \frac{\varepsilon}{S}BB^\top \tag{13}$$

Without explicitly solving this equation, we see that the resulting covariance Σ is proportional to the learning rate ε and inversely proportional to the magnitude of A and minibatch size S. This characterizes the stationary distribution of running SGD with a constant step size.

Constant Stochastic Gradient Descent

Bayesian Inference

In Bayesian inference, we assume a probabilistic model $p(\theta, \boldsymbol{x})$ with data $\boldsymbol{x}$ and hidden variables θ; our goal is to approximate the posterior

$$p(\theta|\boldsymbol{x}) = \exp\{\log p(\theta, \boldsymbol{x}) - \log p(\boldsymbol{x})\} \tag{14}$$

Constant SGD

Assumption 4 says that the posterior is approximately Gaussian in the region that the stationary distribution focuses on

$$f(\theta) \propto \exp\{\frac{N}{2}\theta^\top A\theta\} \tag{15}$$

Consider a more general SGD scheme that may involve a preconditioning matrix H instead of a scalar learning rate ε:

$$\theta_{t+1} = \theta_t - H\hat{g}_S(\theta(t)) \tag{16}$$

Constant Stochastic Gradient Descent

Constant SGD

We will set the parameters of SGD to minimize the KL divergence between the stationary distribution $q(\theta)$ (Eq. 12) and the posterior $f(\theta)$ (Eq. 15).

$$\{H^*, S^*\} = \arg\min_{H,S} KL(q||f) \tag{17}$$

Consider a scalar learning rate ε (or a trivial preconditioner $H = \varepsilon \boldsymbol{I}$)

$$KL(q||f) = -\mathbb{E}_q[\log f(\theta)] + \mathbb{E}_q[\log q(\theta)] =$$

$$= \frac{1}{2}\left(N\mathbb{E}_q[\theta^\top A\theta] - \log|NA| - \log|\Sigma| - D\right) =$$

$$= \frac{1}{2}\left(NTr(A\Sigma) - \log|NA| - \log|\Sigma| - D\right)$$

where $|\cdot|$ is a determinant and D – dimension of θ

Constant Stochastic Gradient Descent

Theorem 1

Theorem 1 (constant SGD) Under Assumptions A1-A4, the constant learning rate that minimizes KL divergence from the stationary distribution of constant SGD to the posterior is

$$\varepsilon^* = 2\frac{S}{N}\frac{D}{Tr(BB^\top)} \tag{18}$$

Proof: Let $\Sigma_0 \equiv \frac{S}{\varepsilon}\Sigma$. According to eq.13 Σ_0 is independent with respect to ε.

$$\log|\Sigma| = D\log\left(\frac{\varepsilon}{S}\right) + \log|\Sigma_0|$$

Since Σ_0 is constant. We also need to simplify the term $Tr(A\Sigma)$

$$Tr(A\Sigma) = \frac{1}{2}(Tr(A\Sigma) + Tr(\Sigma A)) = \frac{\varepsilon}{2S}Tr(BB^\top)$$

Constant Stochastic Gradient Descent

Theorem 1

Proof: The KL divergence is therefore, up to constant terms

$$KL(q||f) \stackrel{c}{=} \frac{\varepsilon N}{2S} Tr(BB^\top) - D \log\left(\frac{\varepsilon}{S}\right) \tag{19}$$

Let $x = \frac{\varepsilon}{S}$, minimizing KL divergence over x gives

$$\frac{\partial KL(q||f)}{\partial x} = \frac{N}{2} Tr(BB^\top) - \frac{D}{x} = 0 \to x = \frac{2}{N} \frac{D}{Tr(BB^\top)}$$

After the reverse substitution, we obtain the required eq. 18.

Constant Stochastic Gradient Descent

Theorem 2

Theorem 2 (Preconditioned constant SGD) The preconditioner for constant SGD that minimizes KL divergence from the stationary distribution to the posterior is

$$H^* = \frac{2S}{N}(BB^\top)^{-1} \tag{20}$$

Proof: Ornstein-Uhlenbeck process which corresponds to preconditioned SGD according to 11

$$d\theta(t) = -HA\theta(t)dt + \frac{1}{\sqrt{S}}HBdW(t)$$

All our results carry over after substituting $A \leftarrow HA$, $\varepsilon B \leftarrow HB$. Then

$$\Sigma A + A\Sigma = \frac{\varepsilon}{S}BB^\top \rightarrow \Sigma HA + HA\Sigma = \frac{1}{S}HBB^\top$$

Constant Stochastic Gradient Descent

Theorem 2

Proof: Eq.13 after the transformation and multiplication by H^{-1} from the left, becomes

$$A\Sigma + H^{-1}\Sigma AH = \frac{1}{S}BB^{\top}H \tag{21}$$

Using the cyclic property of the trace, this implies that

$$Tr(A\Sigma) = \frac{1}{2}(Tr(A\Sigma) + Tr(H^{-1}\Sigma AH)) = \frac{1}{2S}Tr(BB^{\top}H) \tag{22}$$

Define $Q = \Sigma H^{-1}$, hence $Q^{\top} = H^{-1}\Sigma$, since Σ, H and H^{-1} are symmetric. Eq.21 can be written as $QA + AQ^{\top} = \frac{1}{S}BB^{\top}$. Thus, we see that Q is independent of H. The log determinant term is up to a constant $\log|\Sigma| = \log|H| + \log|Q|$.

Constant Stochastic Gradient Descent

Theorem 2

Proof: Combining Eq.22 with this term, the KL divergence is up to a constant

$$KL(q||f) \stackrel{c}{=} \frac{N}{2S} Tr(BB^\top H) + \log|H| + \log|Q| \tag{23}$$

Taking derivatives with respect to the entries of H results in Eq.20.

Corollaries

In high-dimensional applications, working with large dense matrices is impractical. In those settings we can constrain the preconditioner to be diagonal.

Corollary 3: The optimal diagonal preconditioner for SGD that minimizes KL divergence to the posterior is $H^*_{kk} = \frac{2S}{NBB^\top_{kk}}$

Constant SGD as Variational EM

Consider a supervised probabilistic model with joint distribution $p(y, \theta|x, \lambda) = p(y|x, \theta)p(\theta|\lambda)$. Our goal is to find optimal hyperparameters λ. Jointly point-estimating θ and λ by following gradients of the log joint leads to overfitting or degenerate solutions. This can be prevented in a Bayesian approach, where we treat the parameters θ as latent variables.

$$\lambda^* = \arg\max_{\lambda} \log p(y|x, \lambda) = \arg\max_{\lambda} \log \int_{\theta} p(y, \theta|x, \lambda) d\theta$$

Variational expectation maximization tries to find a value for λ that maximizes the expected log-joint probability $\mathbb{E}_q[\log(\theta, y|x, \lambda)]$ where $q(\theta)$ approximate the posterior $p(\theta|y, x, \lambda)$.
Let $\mathcal{L}(\theta, \lambda) = -\log p(\theta, y|x, \lambda)$, then

$$\theta_{t+1} = \theta_t - \varepsilon^* \nabla_\theta \mathcal{L}(\theta_t, \lambda_t), \quad \lambda_{t+1} = \lambda_t - \rho_t \nabla_\lambda \mathcal{L}(\theta_t, \lambda_t) \tag{24}$$

The λ update uses a decreasing learning rate ρ_t and therefore converges to a local optimum.

Stochastic Gradient with Momentum

The updates of SGD with momentum are

$$\begin{cases} v(t+1) = (1-\mu)v(t) - \varepsilon \hat{g}_S(\theta(t)) \\ \theta(t+1) = \theta(t) + v(t+1) \end{cases}$$

As before we assume a quadratic objective $\mathcal{L} = \frac{1}{2}\theta^\top A\theta$. Going through the same steps A1-A4, we find

$$\begin{cases} dv = -\mu v dt - \varepsilon A\theta dt + \frac{1}{\sqrt{S}}\varepsilon B dW \\ d\theta = v dt \end{cases} \tag{25}$$

We solve this set of stochastic equations asymptotically for the long-time limit.

$$d\mathbb{E}[v] = -\mu\mathbb{E}[v]dt - \varepsilon A\mathbb{E}[\theta]dt, \quad d\mathbb{E}[\theta] = \mathbb{E}[v]dt$$

Stochastic Gradient with Momentum

In order to compute the stationary distribution, we derive and solve similar equations for the second moments. We derive the following conditions:

$$\begin{aligned} \mathbb{E}[vv^\top] &= \frac{\varepsilon}{2}\mathbb{E}[\theta\theta^\top]A + \frac{\varepsilon}{2}A\mathbb{E}[\theta\theta^\top], \\ \mu\mathbb{E}[vv^\top] &= \frac{\varepsilon^2}{2S}BB^\top \end{aligned} \tag{26}$$

$\mathbb{E}[vv^\top]$ is a matrix of expected kinetic energies, while $\frac{1}{2}(\mathbb{E}[\theta\theta^\top]A + A\mathbb{E}[\theta\theta^\top])$ has the interpretation of a matrix of expected potential energies. The first equation is the conservation of energy in an equilibrium system. The second equation can be interpreted as the fluctuation-dissipation theorem. Combining both equations and using $\Sigma = \mathbb{E}[vv^\top]$ yields

$$\Sigma A + A\Sigma = \frac{\varepsilon}{\mu S}BB^\top \tag{27}$$

This is exactly Eq.13 of SGD without momentum.

Stochastic Gradient with Momentum

Discussion

The Eq.27 is exactly Eq.13 with the difference that the noise covariance is re-scaled by a factor $\frac{\varepsilon}{\mu S}$ instead of $\frac{\varepsilon}{S}$.
Only the combination $\frac{\varepsilon}{\mu S}$ affects the KL divergence to the posterior. Thus, no single optimal constant learning rate exists—many combinations of ε, μ, and S can yield the same stationary distribution. But different choices of these parameters will affect the dynamics of the Markov chain. For example, Sutskever et al. (2013) observe that, for a given effective learning rate $\frac{\varepsilon}{\mu S}$ using a smaller μ sometimes makes the discretized dynamics of SGD more stable. Also, using very small values of μ while holding $\frac{\varepsilon}{\mu}$ fixed will eventually increase the autocorrelation time of the Markov chain.

Discussion

Estimating noise covariance

In order to use our theoretical insights in practice, we need to estimate the stochastic gradient noise covariance $C \equiv BB^\top$. We do this in an online manner. Let g_t be the full gradient, $\hat{g}_{S,t}$ be the stochastic gradient of the full minibatch and $\hat{g}_{1,t}$ be the stochastic gradient of the first sample in the minibatch at time t. For large S we can approximate $g_t \approx \hat{g}_{S,t}$ and thus obtain an estimator of the noise covariance by $(\hat{g}_{1,t} - \hat{g}_{S,t})(\hat{g}_{1,t} - \hat{g}_{S,t})^\top$ we can now build an online estimate C_t that approaches C by the following recursion

$$C_t = (1 - \kappa_t)C_{t-1} + \kappa_t(\hat{g}_{1,t} - \hat{g}_{S,t})(\hat{g}_{1,t} - \hat{g}_{S,t})^\top \tag{28}$$

κ_t is a decreasing learning rate. Ahn et al. (2012) have proven that such an online average converges to the noise covariance in the optimum at long times (provided that $\kappa_t \sim \frac{1}{t}$ and that N is sufficiently large).

Discussion

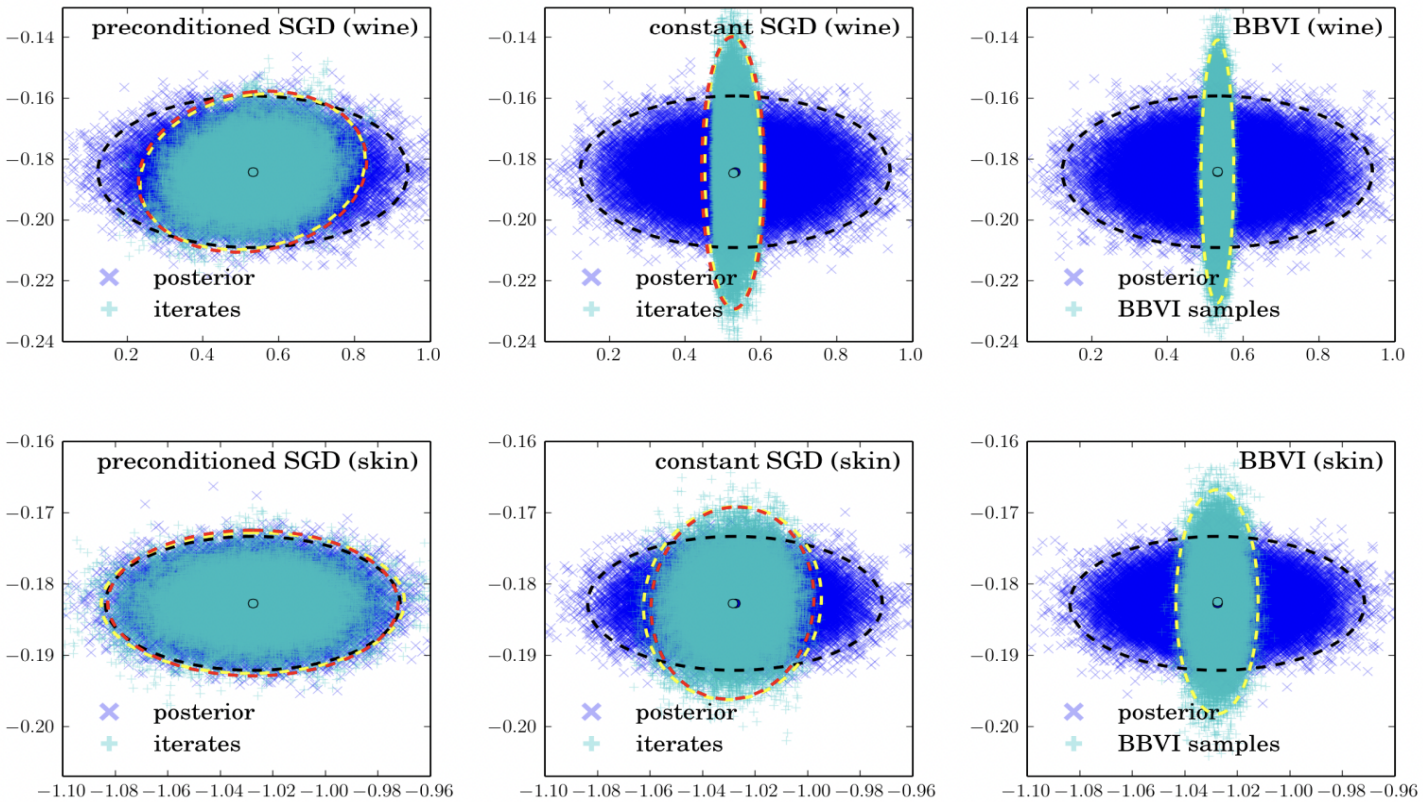

Figure 1: Posterior distribution $f(\theta) \propto \exp\{-N\mathcal{L}(\theta)\}$ (blue) and stationary sampling distributions $q(\theta)$ of the iterates of SGD (cyan) or black box variational inference (BBVI) based on reparameterization gradients. Rows: linear regression (top) and logistic regression (bottom) discussed in Section 7. Columns: full-rank preconditioned constant SGD (left), constant SGD (middle), and BBVI (Kucukelbir et al., 2015) (right). We show projections on the smallest and largest principal component of the posterior. The plot also shows the empirical covariances (3 standard deviations) of the posterior (black), the covariance of the samples (yellow), and their prediction (red) in terms of the Ornstein-Uhlenbeck process, Eq. 13.

Discussion

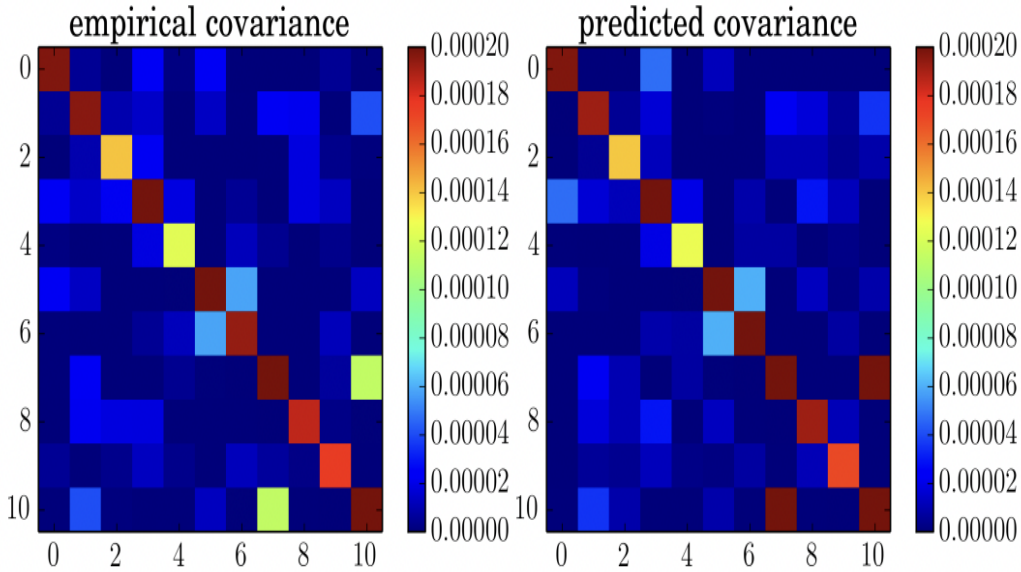

Figure 2: Empirical and predicted covariances of the iterates of stochastic gradient descent, where the prediction is based on Eq. 13. We used linear regression on the wine quality data set as detailed in Section 7.1.