Mathematical Institute

Stochastic gradient descent and its extensions

Theories of Deep Learning: C6.5,
Lecture / Video 7

Prof. Jared Tanner
Mathematical Institute
University of Oxford

Oxford Mathematics

UNIVERSITY OF OXFORD
Mathematical Institute

Consider a fully connected L layer deep net given by

$$h^{(\ell)} = W^{(\ell)} z^{(\ell)} + b^{(\ell)}, \qquad z^{(\ell+1)} = \phi(h^{\ell}), \qquad \ell = 0, \ldots, L-1,$$

for $\ell = 1, \ldots, L$ with nonlinear activation $\phi(\cdot)$ and $W^{(\ell)} \in \mathbb{R}^{n_\ell \times n_\ell}$. The trainable parameters for the DNN, $\theta := \{W^{(\ell)}, b^{(\ell)}\}_{\ell=1}^{L}$ are learned by minimizing a high dimensional, $|\theta| \sim n^2 L$, loss function such as

$$\mathcal{L}(\theta; X, Y) = (2m)^{-1} \sum_{\mu=1}^{m} \sum_{i=1}^{n_L} (H(x_\mu(i); \theta) - y_{i,\mu})^2.$$

The shape of $\mathcal{L}(\theta)$ and our knowledge about a good initial minimizer $\theta^{(0)}$ strongly influence our ability to learn the parameters θ for the DNN.

Gradient calculated through back-propagation

Gradients by passing the error backward through the net

Mathematical Institute

$$\mathcal{L}(\theta; X, Y) = (2m)^{-1} \sum_{\mu=1}^{m} \sum_{i=1}^{n_L} (H(x_\mu(i); \theta) - y_{i,\mu})^2$$

Letting $\delta_\ell := \frac{\partial \mathcal{L}}{\partial h^{(\ell)}}$ and as before $D^{(\ell)}$ the diagonal matrix with $D^{(\ell)}_{ii} = \phi'(h^{(\ell)}_i)$ we have

$$\delta_\ell = D^{\ell}(W^{(\ell)})^T \delta_{\ell+1} \quad \text{and} \quad \delta_L = D^{(L)} \text{grad}_{h^{(L)}} \mathcal{L}.$$

which gives the formula for computing the δ_ℓ for each layer as

$$\delta_\ell = \left(\Pi_{k=\ell}^{L-1} D^{(k)} (W^{(k)})^T \right) D^{(L)} \text{grad}_{h^{(L)}} \mathcal{L}.$$

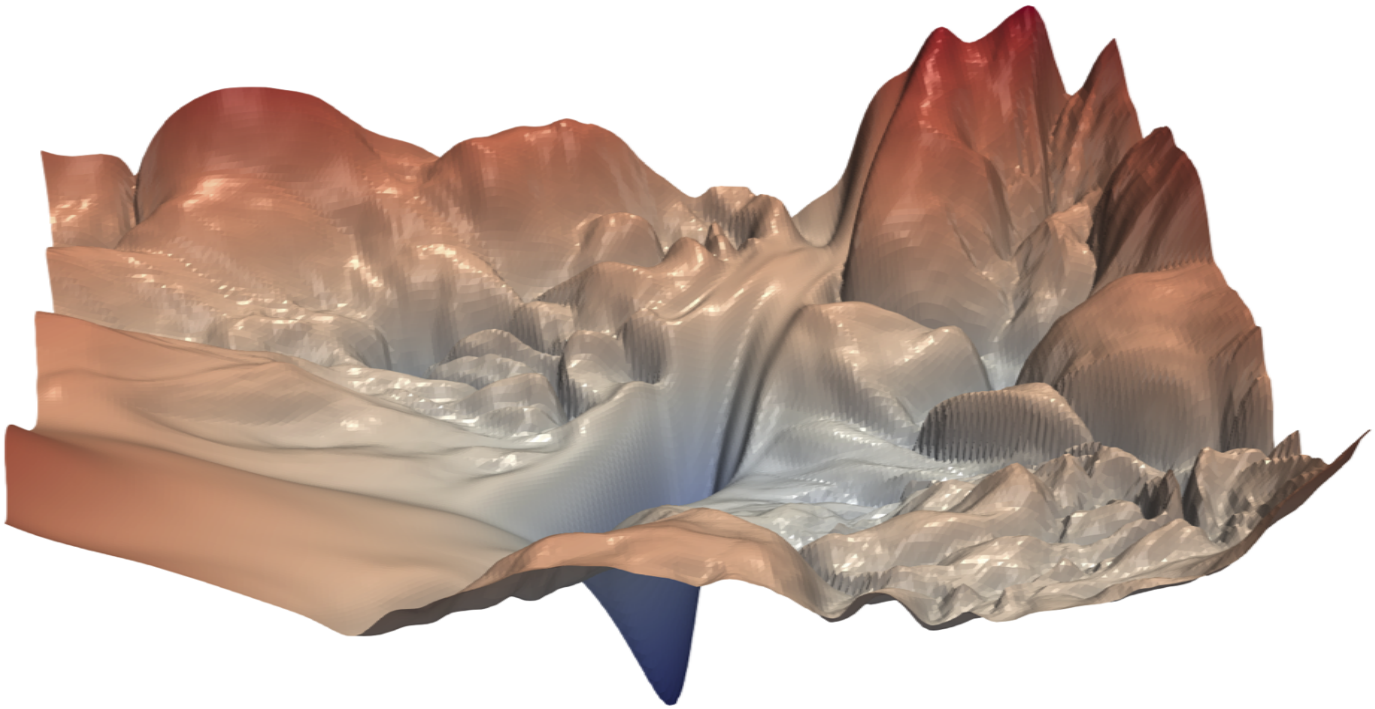
and the resulting gradient $\text{grad}_\theta \mathcal{L}$ with entries as

$$\frac{\partial \mathcal{L}}{\partial W^{(\ell)}} = \delta_{\ell+1} \cdot h_\ell^T \quad \text{and} \quad \frac{\partial \mathcal{L}}{\partial b^{(\ell)}} = \delta_{\ell+1}$$

Loss landscape example: 56 layers fully connected (Li et al. 18')

Loss landscapes of DNNs are typically non-convex

UNIVERSITY OF OXFORD
Mathematical Institute



http://papers.nips.cc/paper/7875-visualizing-the-loss-landscape-of-neural-nets.pdf

Given a loss function $\mathcal{L}(\theta; X, Y)$, gradient descent is given by

$$\theta^{(k+1)} = \theta^{(k)} - \alpha \cdot \text{grad}_\theta \mathcal{L}(\theta, X, Y)$$

with α is referred to as the stepsize, or in DL the "learning rate." In DL $\mathcal{L}(\theta; X, Y)$ is the sum of m individual loss functions for m data point: $\mathcal{L}(\theta; X, Y) = m^{-1} \sum_{\mu=1}^{m} l(\theta; x_\mu, y_\mu)$
For $m \gg 1$ gradient descent is computationally too costly and instead one can break appart the m loss functions into "mini-batches" and repeatedly solve

$$\theta^{(k+1)} = \theta^{(k)} - \alpha |S_k|^{-1} \text{grad}_\theta \sum_{\mu \in S_k} l(\theta; x_\mu, y_\mu).$$

This is referred to as stochastic gradient descent as typically S_k is chosen in some randomized method, usually as a partition of $[m]$ and a sequence of S_k which cover $[m]$ is referred to as an "epoch."

- SGD is preferable for large m as it reduces the per iteration computational cost dependence on m to instead depend on $|S_k|$ which can be set by the user as opposed to m which is given by the data set.
- SGD, and gradient descent, require selection of a learning rate (stepsize) which in deep learning is typically selected using some costly trial and error heuristics.
- The learning rate is typically chosen adaptively in a way that satisfies $\sum_{k=1}^{\infty} \alpha_k = \infty$ and $\sum_{k=1}^{\infty} \alpha_k^2 < \infty$; in particular as $\alpha_k \sim k^{-1}$.
- The optimal selection of learning weight, and selection of S_k, depends on the unknown local Lipschitz constant $\|\text{grad}I(\theta_1; x_\mu, y_\mu) - \text{grad}I(\theta_2; x_\mu, y_\mu)\| \leq L_\mu \|\theta_1 - \theta_2\|$.

Lemma 1 [An overestimation property] Let $\mathcal{L}(\theta) \in C^1(\mathbb{R}^n)$ with $\nabla\mathcal{L}$ Lipshitz continuous with constant L. Then for any θ and $d \in \mathbb{R}^n$ and $\alpha \in \mathbb{R}$:

$$\mathcal{L}(\theta + \alpha d) \leq \mathcal{L}(\theta) + \alpha \nabla\mathcal{L}(\theta)^T d + \alpha^2 \frac{L}{2}\|d\|^2.$$

In particular, if $d = -\nabla\mathcal{L}(\theta)$ then

$\mathcal{L}(\theta - \alpha\nabla\mathcal{L}(\theta)) \leq \mathcal{L}(\theta) - \alpha\|\nabla\mathcal{L}(\theta)\|^2 + \frac{L}{2}\alpha^2\|\nabla\mathcal{L}(\theta)\|_2^2$ and so

$$\mathcal{L}(\theta - \alpha\nabla\mathcal{L}(\theta)) \leq \mathcal{L}(\theta) - \alpha\left(1 - \frac{L}{2}\alpha\right)\|\nabla\mathcal{L}(\theta)\|^2 \quad (OP_{GD}).$$

Proof of Lemma 1. By Taylor's theorem in integral form we have

$$\begin{aligned}
&\mathcal{L}(\theta+\alpha d) = \mathcal{L}(\theta) + \int_{t=0}^{t=1} \nabla\mathcal{L}(\theta+\alpha t d)^T(\alpha d)\cdot dt \\
&= \mathcal{L}(\theta) + \alpha\nabla\mathcal{L}(\theta)^T d + \alpha\int_{t=0}^{t=1}[\nabla\mathcal{L}(\theta+\alpha t d) - \nabla\mathcal{L}(\theta)]^T d\cdot dt \\
&\le \mathcal{L}(\theta) + \alpha\nabla\mathcal{L}(\theta)^T d + \alpha\int_{t=0}^{t=1}\|\nabla\mathcal{L}(\theta+\alpha t d) - \nabla\mathcal{L}(\theta)\|\cdot\|d\| dt \\
&\qquad\qquad \text{by Cauchy-Schwarz inequality} \\
&\le \mathcal{L}(\theta) + \alpha\nabla\mathcal{L}(\theta)^T d + \alpha L\|d\|\int_{t=0}^{t=1}\|\theta+\alpha t d-\theta\| dt \\
&\qquad\qquad \text{by Lipschitz continuity of the gradient} \\
&\le \mathcal{L}(\theta) + \alpha\nabla\mathcal{L}(\theta)^T d + \alpha^2 L\|d\|^2\int_{t=0}^{t=1} t dt,
\end{aligned}$$

which gives the required overestimation inequality.□

If $|\mathcal{S}_k| = 1$ (one data element), the expected gradient wrt data point $G^k := \text{grad}_\theta \sum_{\mu \in S_k} l(\theta; x_\mu, y_\mu)$:

$$\mathrm{E}_{\mathcal{S}_k}[G^k] = \mathrm{E}[G^k|\mathcal{S}_k] = \sum_{i=1}^{m} \mathrm{E}[G^k|\mathcal{S}_k = i] \cdot \mathrm{P}[\mathcal{S}_k = i] = \sum_{i=1}^{m} \nabla l_i(\theta^k) \cdot \frac{1}{m} = \nabla \mathcal{L}(\theta^k).$$

- Similarly for larger sets $\mathcal{S}_k$ drawn uniformly from $\binom{m}{|\mathcal{S}_k|}$ possible configurations; referred to as mini-batches.
- Above, we used $\mathrm{E}[G^k|\mathcal{S}_k = i] = \nabla l_i(\theta^k)$ (true due to iid choice of $\mathcal{S}_k$ and G^k). More generally, we require an unbiased estimator of the true gradient: $\mathrm{E}_{\mathcal{S}_k}[G^k] = \nabla \mathcal{L}(\theta^k)$.
- (A realization of) $-G^k$ may not be a descent direction: $\nabla \mathcal{L}(\theta^k)^\mathsf{T}(-G^k) < 0$ cannot be guaranteed, but is guaranteed in expectation. Therefore, we analyse the expected descent of the random iterates (θ^k).

Global convergence of SGD - in expectation

Mathematical Institute

Assumptions for our analysis ($|\mathcal{S}_k| = 1$):

(1) for all $i \leq m$, ∇l_i is Lipschitz continuous, constant L
$\Rightarrow \nabla l$ Lipschitz continuous, L

(2) $\exists\, M > 0$ s.t.
$\mathrm{VAR}(G^k|\mathcal{S}_k) := \mathrm{E}[(G^k - \nabla l(\theta^k))^{\mathsf{T}}(G^k - \nabla l(\theta^k))|\mathcal{S}_k] \leq M$ for all k
(bounded total variance can usually be guaranteed in a neighbourhood of θ^* but not globally for strongly convex $\mathcal{L}(\cdot)$.)

Recall that G^k conditioned on current batch is an unbiased estimator of the true gradient; this is true here (and when $|\mathcal{S}_k| > 1$), but it would have to be assumed in a more general stochastic framework. (A more thorough analysis would also condition on θ^k.)

Lemma 2 [An overestimation property - in expectation] Assume Assumption (1) holds. When applying SGD to $\mathcal{L}(\theta)$ with $|\mathcal{S}_k| = 1$, we have

$$\mathrm{E}_{\mathcal{S}_k}\left[\mathcal{L}(\theta^{k+1})\right] \leq \mathcal{L}(\theta^k) - \alpha \nabla \mathcal{L}(\theta^k)^{\mathsf{T}} \mathrm{E}_{\mathcal{S}_k}\left[G^k\right] + \frac{L\alpha^2}{2} \mathrm{E}_{\mathcal{S}_k}\left[\|G^k\|^2\right].$$

If Assumption (2) also holds, then

$$\mathrm{E}_{\mathcal{S}_k}\left[\mathcal{L}(\theta^{k+1})\right] \leq \mathcal{L}(\theta^k) - \alpha^k \left(\frac{L\alpha^k}{2} - 1\right) \|\nabla \mathcal{L}(\theta^k)\|^2 + \frac{ML(\alpha^k)^2}{2}.$$

Global convergence of SGD - in expectation

Mathematical Institute

Proof of Lemma 2. Apply Lemma 1 to $\mathcal{L}$ with $\theta = \theta^k$, $d = G^k$ and $\alpha = \alpha^k$: using $\theta^{k+1} = \theta^k + \alpha^k G^k$,

$$\mathcal{L}(\theta^{k+1}) \leq \mathcal{L}(\theta^k) - \alpha^k \nabla \mathcal{L}(\theta^k)^T G^k + \frac{L}{2}(\alpha^k)^2 \|G^k\|^2.$$

Applying expectation on both sides wrt $\mathcal{S}_k$,

$$\mathrm{E}_{\mathcal{S}_k}[\mathcal{L}(\theta^{k+1}] \leq \mathcal{L}(\theta^k) - \alpha^k \nabla \mathcal{L}(\theta^k)^T \mathrm{E}_{\mathcal{S}_k}[G^k] + \frac{L}{2}(\alpha^k)^2 \mathrm{E}_{\mathcal{S}_k}[\|G^k\|^2].$$

where we used that $\mathcal{L}(\theta^k)$ and $\nabla \mathcal{L}(\theta^k)$ do not depend on $\mathcal{S}_k$. We already showed/assumed that $\mathrm{E}_{\mathcal{S}_k}[G^k] = \nabla \mathcal{L}(\theta^k)$.

$$\begin{aligned} \mathrm{VAR}(G^k|\mathcal{S}_k) &= \mathrm{E}_{\mathcal{S}_k}\left[\|G^k\|^2\right] - 2\nabla \mathcal{L}(\theta^k)^{\mathsf{T}} \mathrm{E}_{\mathcal{S}_k}[G^k] + \|\nabla \mathcal{L}(\theta^k)\|^2 \\ &= \mathrm{E}_{\mathcal{S}_k}\left[\|G^k\|^2\right] - \|\nabla \mathcal{L}(\theta^k)\|^2. \end{aligned}$$

Thus $\mathrm{E}_{\mathcal{S}_k}\left[\|G^k\|^2\right] \leq M + \|\nabla \mathcal{L}(\theta^k)\|^2$. $\square$

Global convergence of SGD: the strongly convex case

Let $\mathcal{L}$ be (for now) strongly convex with parameter $\mu > 0$, namely $\mathcal{L}(x+s) \geq \mathcal{L}(x) + s^T \nabla \mathcal{L}(x) + \frac{\mu}{2}\|s\|^2$ for all x, s.

Theorem 3 Let $\mathcal{L}$ be smooth, strongly convex and satisfying Assumption (1), (2). Let SGD with fixed stepsize be applied to minimize $\mathcal{L}$, where $\alpha^k = \underline{\alpha} = \frac{\eta}{L}$ where $\eta \in (0, 1]$. Then SGD converges linearly to a residual error in the following sense: for all $k \geq 0$,

$$\mathrm{E}[\mathcal{L}(\theta^k)] - \mathcal{L}(\theta^*) - \frac{\eta M}{2\mu} \leq \left(1 - \frac{\eta\mu}{L}\right)^k \cdot \left[\mathcal{L}(\theta^0) - f(\theta^*) - \frac{\eta M}{2\mu}\right].$$

- Thus $\lim_{k\to\infty}(\mathrm{E}[\mathcal{L}(\theta^k)] - \mathcal{L}(\theta^*)) \leq \frac{\underline{\alpha} ML}{2\mu} = \frac{\eta M}{2\mu}$. Convergence is obtained, in expectation, up to the level $\frac{\eta M}{2\mu}$ (noise level !), which can be decreased in various ways.
- The ratio $\frac{L}{\mu}$ is a condition number of $\mathcal{L}$ (connect to second derivatives).

Global convergence of SGD: the strongly convex case

Mathematical Institute

Proof of Theorem 3. Lemma 3 and $\frac{L\alpha}{2} - 1 = \frac{\eta}{2} - 1 < -\frac{1}{2}$ give

$$\mathsf{E}_{\mathcal{S}_k}\left[\mathcal{L}(\theta^{k+1})\right] \leq \mathcal{L}(\theta^k) - \frac{\alpha}{2}\|\nabla\mathcal{L}(\theta^k)\|^2 + \frac{ML\alpha^2}{2}.$$

Taking expectation E with respect to the past, namely, $\mathcal{S}_0, \ldots, \mathcal{S}_{k-1}$ on both sides of the above, we note that we have a memoryless property so current iterate only depends on previous sample size ($\mathsf{E} = \mathsf{E}_k := \mathsf{E}(\cdot|\mathcal{S}_0, \ldots, \mathcal{S}_k) = \mathsf{E}_{\mathcal{S}_k}$):

$$\mathsf{E}_k\left[\mathcal{L}(\theta^{k+1})\right] - \mathcal{L}(\theta^*) \leq \mathsf{E}_{k-1}\left[\mathcal{L}(\theta^k)\right] - \mathcal{L}(\theta^*) - \frac{\alpha}{2}\mathsf{E}_{k-1}\left[\|\nabla\mathcal{L}(\theta^k)\|^2\right] + \frac{ML\alpha^2}{2}.$$

A consequence the strong convexity property, is that θ^* global minimizer is unique and $\mathcal{L}(\theta^k) - \mathcal{L}(\theta^*) \geq \frac{1}{2\mu}\|\nabla\mathcal{L}(\theta^k)\|^2$; thus

$$\mathsf{E}_{k-1}(\mathcal{L}[(\theta^k] - \mathcal{L}(\theta^*)) \geq \frac{1}{2\mu}\mathsf{E}_{k-1}(\|\nabla\mathcal{L}(\theta^k)\|^2).$$

Proof of Theorem 3. (continued) We deduce

$$\mathsf{E}_k\left[\mathcal{L}(\theta^{k+1})\right]-\mathcal{L}(\theta^*) \leq (1-\mu\alpha)\left(\mathsf{E}_{k-1}\left[\mathcal{L}(\theta^k)\right]-\mathcal{L}(\theta^*)\right)+\frac{ML\alpha^2}{2},$$

or equivalently,

$$\mathsf{E}_k\left[\mathcal{L}(\theta^{k+1})\right]-\mathcal{L}(\theta^*)-\frac{\alpha ML}{2\mu} \leq (1-\mu\alpha))\left(\mathsf{E}_{k-1}\left[\mathcal{L}(\theta^k)\right]-\mathcal{L}(\theta^*)-\frac{\alpha ML}{2\mu}\right).$$

Note that $\alpha = \eta/L \leq 1/L \leq 1/\mu$. Replacing α gives

$$\mathsf{E}_k\left[\mathcal{L}(\theta^{k+1})\right]-\mathcal{L}(\theta^*)-\frac{M\eta}{2\mu} \leq \left(1-\frac{\eta\mu}{L}\right)\left(\mathsf{E}_{k-1}\left[\mathcal{L}(\theta^k)\right]-\mathcal{L}(\theta^*)-\frac{M\eta}{2\mu}\right),$$

The claim now follows by induction.□

Decreasing the SGD "noise floor": technique 1

Though not always desirable (due to the needs for small 'generalization error'), the SGD "floor" (noise level) of $\frac{\eta M}{2\mu}$ can be removed so that $\lim_{k\to\infty} \mathsf{E}[\mathcal{L}(\theta^k)] = \mathcal{L}(\theta^*)$.

Dynamic stepsize reduction. Technique 1: Dynamically reduce $\alpha^k = \frac{\eta_k}{L}$. Note that $\eta_k \to 0$ makes the residual $\frac{\eta_k M}{2\mu} \to 0$ but it also means that $\left(1 - \frac{\eta_k}{L}\right) \to 1$, so the price is that we lose linear convergence!

Theorem 4. [Dynamic stepsize stochastic gradient descent (DS-SGD)] Let $\alpha^k = \frac{2}{2L+k\mu}$, for all $k \geq 0$. Then SGD satisfies

$$0 \leq \mathsf{E}[\mathcal{L}(\theta^k)] - \mathcal{L}(\theta^*) \leq \frac{\nu}{2\frac{L}{\mu} + k} \qquad (*)$$

for all $k \geq 0$, where $\nu := 2\frac{L}{\mu} \times \max\left\{\frac{M}{\mu}, \mathcal{L}(\theta^0) - \mathcal{L}(\theta^*)\right\}$. Thus $\lim_{k\to\infty} \mathsf{E}[\mathcal{L}(\theta^k)] = \mathcal{L}(\theta^*)$. But rate is $\mathcal{O}\left(\frac{1}{k}\right)$ - sublinear !

Dynamic stepsize reduction (continued)

Proof of Theorem 4. (similar to proof of Theorem 3) Note that $\alpha^k \leq 1/L \leq 1/\mu$ and all arguments continue to hold in the proof of Th 3 until and including, and so for all $k \geq 0$,

$$\mathsf{E}_k\left[\mathcal{L}(\theta^{k+1})\right] - \mathcal{L}(\theta^*) - \frac{\alpha^k ML}{2\mu} \leq (1-\mu\alpha^k)\left(\mathsf{E}_{k-1}\left[\mathcal{L}(\theta^k)\right] - \mathcal{L}(\theta^*) - \frac{\alpha^k ML}{2\mu}\right).$$

We are now going to prove the desired conclusion (*) by induction. Clearly at $k=0$, (*) holds. Assume (*) holds at $k>0$, and substitute (*) into the above displayed equation. We obtain

$$\mathsf{E}_k\left[\mathcal{L}(\theta^{k+1})\right] - \mathcal{L}(\theta^*) - \frac{\alpha^k ML}{2\mu} \leq (1-\mu\alpha^k)\left(\frac{\nu}{2\frac{L}{\mu}+k} - \frac{\alpha^k ML}{2\mu}\right).$$

Using the expression of α^k in the above and simplifying the expressions provides (*) with k replaced by $(k+1)$.$\square$

Increase mini-batch sizes from $|\mathcal{S}_k| = 1$ to $|\mathcal{S}_k| = p \geq 1$. Use $G^k = \frac{1}{p}\sum_{j\in\mathcal{S}_k} \nabla l_j(\theta^k)$, where $j \in \mathcal{S}_k$ i.i.d. $\sim \mathcal{U}(\{1,\ldots,m\})$:

$$\begin{aligned}\mathrm{VAR}(G^k|\mathcal{S}_k) &= \sum_{j\in\mathcal{S}_k} \frac{1}{p^2} \mathrm{E}_{\mathcal{S}_k}\left[\|\nabla l_j(\theta^k) - \nabla\mathcal{L}(\theta^k)\|^2\right] \\ &\quad +2\sum_{j<i} \frac{1}{p^2} \mathrm{E}_{\mathcal{S}_k}\left[\nabla l_j(\theta^k) - \nabla\mathcal{L}(\theta^k)\right]^{\mathsf{T}} \mathrm{E}_{\mathcal{S}_k}\left[\nabla l_i(\theta^k) - \nabla\mathcal{L}(\theta^k)\right] \\ &= \frac{1}{p^2}\sum_{j\in\mathcal{S}_k} \mathrm{VAR}(\nabla l_j(\theta^k)) + 0 \leq \frac{M}{p},\end{aligned}$$

where we have used $|\mathcal{S}_k| = p$ and the independence of i and j indices in $\mathcal{S}_k$ in the first equality as well as the lack of bias $\mathrm{E}_{\mathcal{S}_k}\left[\nabla l_j(\theta^k)\right] = \nabla\mathcal{L}(\theta^k)$. We also have $\mathrm{E}_{\mathcal{S}_k}\left[G^k\right] = \nabla\mathcal{L}(\theta^k)$ - unbiased batch gradient.

Decreasing the SGD "noise floor": technique 2

Increase mini-batch sizes from $|\mathcal{S}_k| = 1$ to $|\mathcal{S}_k| = p \geq 1$.
(continued)
Then, as in Theorem 3, we deduce, under the same assumptions,

$$\mathrm{E}[\mathcal{L}(\theta^k)] - \mathcal{L}(\theta^*) - \frac{\eta M}{2\mu p} \leq \left(1 - \frac{\eta\mu}{L}\right)^k \cdot \left[\mathcal{L}(\theta^0) - \mathcal{L}(\theta^*) - \frac{\eta M}{2\mu p}\right].$$

Thus the noise level is decreased by batch size p, without impacting the convergence factor.
(Compare and contrast Techniques 1 and 2.)

Decreasing the SGD "noise floor": technique 3

UNIVERSITY OF OXFORD
Mathematical Institute

Momentum for gradient variance reduction

Technique 3: use acceleration by momentum to reduce VAR($G^k|\mathcal{S}_k$). This yields $E[\mathcal{L}(\theta^k)] \to \mathcal{L}(\theta^*)$ with linear convergence rate, with a much smaller cost per iteration than mini-batching (see the 'Katyusha' paper). `https://www.jmlr.org/papers/volume18/16-410/16-410.pdf`

Other techniques (earlier than Katyusha): variance reduction (SVRG), SAG (Schmidt, Le Roux, Bach'15: restores linear rate for SGD), SAGA (Defazio et al'14).

Conclusions: each of the three approaches for accelerating SGD have merit and are often all used at once. In particular, once SGD appears to stagnate one both reduces the stepsize and increases the batch-size; though this is stopped once validation error begins to increase.

What about SGD performance when $\mathcal{L}$ is nonconvex (as in DNNs)?

Theorem 5. [SGD with fixed stepsize] Let $\mathcal{L} \in \mathcal{C}^1(\mathbb{R}^n)$ be bounded below by $\mathcal{L}_{\text{low}}$, with $\nabla\mathcal{L}$ Lipschitz continuous with Lipschitz constant L (Assumption (1)). Let Assumption (2) hold (bounded variance). Apply the SGD method with fixed stepsize $\alpha = \eta/L$ and $|\mathcal{S}|_k = 1$, where $\eta \in (0, 1]$, to minimizing $\mathcal{L}$. Then

$$\min_{0\leq i\leq k} \mathsf{E}[\|\nabla\mathcal{L}(\theta^i)\|^2] \leq \alpha LM + \frac{2(\mathcal{L}(\theta^0) - \mathcal{L}_{\text{low}})}{k\alpha} = \eta M + \frac{2L(\mathcal{L}(\theta^0) - \mathcal{L}_{\text{low}})}{k\eta}$$

and so the SGD method takes at most $k \leq \frac{2L(\mathcal{L}(\theta^0)-\mathcal{L}_{\text{low}})}{\eta\epsilon}$ iterations/evaluations to generate $\mathsf{E}[\|\nabla\mathcal{L}(\theta^k)\|^2] \leq \epsilon + \eta M$.

- again, note the 'noise floor' that limits the accuracy that can be obtained.

Proof of Theorem 5. The first part of Theorem 3 still applies, and we still have the following expected decrease:

$$\mathsf{E}_k\left[\mathcal{L}(\theta^{k+1})\right] \leq \mathsf{E}_{k-1}\left[\mathcal{L}(\theta^k)\right] - \frac{\alpha}{2}\mathsf{E}_{k-1}\left[\|\nabla\mathcal{L}(\theta^k)\|^2\right] + \frac{ML\alpha^2}{2}.$$

We need to connect the per iteration decrease with the gradient. We have for all $k \geq 0$:

$$\mathsf{E}_{k-1}\left[\mathcal{L}(\theta^k)\right] - \mathsf{E}_k\left[\mathcal{L}(\theta^{k+1})\right] \geq \frac{\alpha}{2}\mathsf{E}_{k-1}\left[\|\nabla\mathcal{L}(\theta^k)\|^2\right] - \frac{ML\alpha^2}{2}.$$

Summing up the above bound from $i = 0$ to k, we deduce

$$\begin{aligned}
\mathcal{L}(\theta^0) - \mathcal{L}_{\text{low}} &\geq \mathcal{L}(\theta^0) - \mathsf{E}_k\left[\mathcal{L}(\theta^{k+1})\right] \\
&\geq \tfrac{\alpha}{2}\textstyle\sum_{i=0}^{k}\mathsf{E}_{i-1}\left[\|\nabla\mathcal{L}(\theta^i)\|^2\right] - (k+1)\tfrac{ML\alpha^2}{2}. \\
&\geq \tfrac{\alpha}{2}(k+1)\left[\min_{0\leq i\leq k}\mathsf{E}[\|\nabla\mathcal{L}(\theta^i)\|^2] - ML\alpha\right]
\end{aligned}$$

□

To reduce the 'noise floor' use: decreasing stepsize, mini-batching. (Acceleration/momentum difficult in the nonconvex case.)

Re decreasing stepsize, let $\alpha^k = \eta_k/L$ where $\eta_k \in (0, 1]$.

Similarly to the proof of Theorem 5, we obtain

$$\sum_{i=0}^{k} \alpha^i \, \mathsf{E}_{i-1} \left[\|\nabla \mathcal{L}(\theta^i)\|^2 \right] \leq 2(\mathcal{L}(\theta^0) - \mathcal{L}_{\text{low}}) + ML \sum_{i=0}^{k} (\alpha^i)^2.$$

And so to reduce the noise term, assume that $\sum_{i=0}^{\infty} \alpha^i = \infty$ and $\sum_{i=0}^{\infty} (\alpha^i)^2 < \infty$.