CPSC536C - UBC, 2026

Student Lecture 1: Analysing stochastic gradient descent

Lecturer: Kevin Thomas *Scribe: Arqam Rizwan*

1 Motivation

Consider the problem of estimating the parameters x of a linear regression model using a least squares loss. This corresponds to minimising a function of the form:

$$f(x) = \frac{1}{2n}\|Ax - b\|^2 = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{2}(\langle a_i, x\rangle - b_i)^2 := \frac{1}{n}\sum_{i=1}^{n} f_i(x)$$

To carry out gradient descent, we would need to evaluate the gradient which is:

$$\nabla f(x) = \frac{1}{n}A^\top(Ax - b) = \frac{1}{n}\sum_{i\leq n}\nabla f_i(x)$$

This is expensive to compute when the number of data points n is large. The idea of stochastic gradient descent is to use a cheaper, unbiased estimator of the gradient- for example using just one random datapoint every iteration.

If $i \sim$Unif$(1, 2, \ldots, n)$:

$$\mathbb{E}[\nabla f_i(x)] = \frac{1}{n}\sum_{j}\nabla f_j(x) = \nabla f(x)$$

In stochastic gradient descent, at each iteration, we sample $i_t \sim$Unif$\{1, 2, \ldots, n\}$ and then update $x_{t+1} := x_t - \eta\nabla f_{i_t}(x)$.

2 Alternative analysis of gradient descent

We want to find: $\min_x f(x)$ where f is a β-smooth convex function. In standard gradient descent, we iterate:

$$x_{t+1} := x_t - \tau\nabla f(x)$$

Let x^* be the optimal point such that for all y, $f(x^*) \leq f(y)$

Lemma 2.1. *If f is β-smooth then $\|\nabla f(x_t)\|^2 \leq 2\beta[f(x_t) - f(x^*)]$*

Proof. By the smoothness assumption, for all y we have

$$\begin{aligned} f(y) &\leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\beta}{2}\|y - x\|^2 \\ f\left(x_t - \frac{1}{\beta}\nabla f(x_t)\right) &\leq f(x_t) - \frac{1}{\beta}\|\nabla f(x_t)\|^2 + \frac{1}{2\beta}\|\nabla f(x_t)\|^2 \\ \frac{1}{2\beta}\|\nabla f(x_t)\|^2 &\leq f(x_t) - f\left(x_t - \frac{1}{\beta}\nabla f(x_t)\right) \\ &\leq f(x_t) - f(x^*) \qquad \because -f(y) \leq -f(x^*) \end{aligned}$$

□

Theorem 2.2. *Let f be convex, β-smooth and assume $\|x_0 - x^*\|_2 \leq R$. Then for update*

$$x_{t+1} := x_t - \frac{\nabla f(x_t)}{2\beta}$$

the function gap can be bounded as

$$\min_{t<T} f(x_t) - f(x^*) \leq \frac{2\beta R^2}{T}$$

i.e. we achieve an ϵ-relative function gap in $O\left(\frac{1}{\epsilon}\right)$ iterations.

Proof. Recall the first order definition of convexity, $f(y) - f(x) \geq \langle y - x, \nabla f(x) \rangle$

$$\begin{aligned} \|x_{t+1} - x^*\|^2 &= \|x_t - \tau\nabla f(x_t) - x^*\|^2 \\ &= \|x^* - x_t\|^2 + \tau^2\|\nabla f(x_t)\|^2 + 2\tau\langle x^* - x_t, \nabla f(x_t)\rangle \\ &\leq \|x_t - x^*\|^2 + \tau^2\|\nabla f(x_t)\|^2 + 2\tau[f(x^*) - f(x_t)] \qquad \text{convexity} \\ &\leq \|x_t - x^*\|^2 + 2\tau^2\beta\left[f(x_t) - f(x^*)\right] - 2\tau[f(x_t) - f(x^*)] \qquad \text{lemma} \\ 2\tau(1 - \tau\beta)\left[f(x_t) - f(x^*)\right] &\leq \|x_t - x^*\|^2 - \|x_{t+1} - x^*\|^2 \end{aligned}$$

Choosing τ such that $1 - \tau\beta = \frac{1}{2} \iff \tau = \frac{1}{2\beta}$, we get

$$\begin{aligned} \frac{1}{2\beta}\left[f(x_t) - f(x^*)\right] &\leq \|x_t - x^*\|^2 - \|x_{t+1} - x^*\|^2 \\ \sum_{t=0}^{T-1}\frac{1}{2\beta}\left[f(x_t) - f(x^*)\right] &\leq \sum_{t=0}^{T-1}\left(\|x_t - x^*\|^2 - \|x_{t+1} - x^*\|^2\right) \\ \frac{1}{2\beta}\sum_{t=0}^{T-1}\left[f(x_t) - f(x^*)\right] &\leq \|x_0 - x^*\|^2 - \|x_T - x^*\|^2 \\ \min_{t<T} f(x_t) - f(x^*) &\leq \frac{1}{T}\sum_{t=0}^{T-1}\left[f(x_t) - f(x^*)\right] \\ &\leq \frac{2\beta}{T}\|x_0 - x^*\|^2 \end{aligned}$$

□

3 Stochastic gradient descent for smooth convex functions

In general, if we have an unbiased estimator of the gradient that is cheaper to compute, then even noisy evaluations of the gradient allow us to move closer to the optimum on average. So we have a trade off- slower convergence due to variance in the gradient estimator for faster updates.

3.1 Assumptions

- Convexity: f is convex
- Smoothness: f is β-smooth.
- Unbiased gradient estimator $\mathbb{E}[\hat{\nabla} f(x)] = \nabla f(x)$
- Bounded variance of gradient estimator: Var$[\hat{\nabla} f(x)] < \sigma^2$

Theorem 3.1. *Let the above assumptions hold and assume $\|x_0 - x^*\|_2 \leq R$. Then for update*

$$x_{t+1} := x_t - \frac{\hat{\nabla} f(x_t)}{2\beta\sqrt{t}}$$

for $t = 0, 1, \ldots, T$ the expected function gap scales as

$$\mathbb{E}[f(\bar{x}_T) - f(x^*)] \leq \frac{4\beta^2 R^2 + \sigma^2(1 + \log T)}{4\beta(\sqrt{T} - 1)} = O\left(\frac{\log T}{\sqrt{T}}\right)$$

where

$$\tau_t = \frac{1}{\sqrt{t}}, \bar{x}_T = \frac{\sum_{t=0}^{T} \tau_t x_t}{\sum_{t=0}^{T} \tau_t}$$

Proof. We use the fact that β-smoothness is equivalent to β-Lipschitzness of the gradient for a convex differentiable function.

$$\begin{aligned}
\|\nabla f(x) - \nabla f(y)\| &= \|\mathbb{E}[\hat{\nabla} f(x)] - \mathbb{E}[\hat{\nabla} f(y)]\| \\
&= \|\mathbb{E}[\hat{\nabla} f(x) - \hat{\nabla} f(y)\| \\
&\leq \mathbb{E}\|\hat{\nabla} f(x) - \hat{\nabla} f(y)\| \qquad \text{Jensen's inequality} \\
&\leq \mathbb{E}[\beta_i \|x - y\|] \qquad \text{smoothness} \\
&\leq \beta\|x - y\|
\end{aligned}$$

Define the conditional expectation given the past iterates at time t, $\mathbb{E}_t[\cdot] := \mathbb{E}[\cdot|x_0, \ldots, x_t]$.

$$\begin{aligned}
\|x_{t+1} - x^*\|^2 &= \|x_t - \tau_t \hat{\nabla} f(x_t) - x^*\|^2 \\
&= \|x_t - x^*\|^2 + 2\tau_t \langle \hat{\nabla} f(x_t), x^* - x_t \rangle + \tau_t^2 \|\hat{\nabla} f(x_t)\|^2 \\
\mathbb{E}[\|x_{t+1} - x^*\|^2 | x_0, \ldots x_t] &= \|x_t - x^*\|^2 + 2\tau_t \mathbb{E}_t[\langle \hat{\nabla} f(x_t), x^* - x_t \rangle] + \tau_t^2 \mathbb{E}_t \|\hat{\nabla} f(x_t)\|^2 \\
&= \|x_t - x^*\|^2 + 2\tau_t \langle \nabla f(x_t), x^* - x_t \rangle + \tau_t^2 \mathbb{E}_t \|\hat{\nabla} f(x_t)\|^2 \\
&\leq \|x_t - x^*\|^2 + 2\tau_t [f(x^*) - f(x_t)] + \tau_t^2 \mathbb{E}_t \|\hat{\nabla} f(x_t)\|^2 \qquad \text{convexity}
\end{aligned}$$

There's a bunch of different assumptions we could use to control the final term. The simplest would be to assume the second moment of the gradient norm is bounded i.e. $\mathbb{E}\|\hat{\nabla} f(x_t)\|^2 \leq \sigma^2$ (ie both the mean and variance are bounded), but this is quite strong. An even stronger alternative assumption would be to assume a Lipschitz condition on each f_i, which would bound the gradient norm directly. Instead, we assume that the gradient estimator has finite variance and is β-smooth.

$$\begin{aligned}
\text{Var}[\hat{\nabla} f(x_t)] &= \mathbb{E}\|\hat{\nabla} f(x_t) - \nabla f(x_t)\|^2 \leq \sigma^2 \\
\mathbb{E}_t\|\hat{\nabla} f(x_t)\|^2 &= \mathbb{E}_t\|\hat{\nabla} f(x_t) - \nabla f(x_t)\|^2 + \mathbb{E}_t\|\nabla f(x_t)\|^2 + 2\mathbb{E}_t\langle \hat{\nabla} f(x_t) - \nabla f(x_t), \nabla f(x_t) \rangle \\
&\leq \sigma^2 + 2\beta[f(x_t) - f(x^*)] + 0 \qquad \text{using Lemma 2.1 and unbiasedness} \qquad (1)
\end{aligned}$$

Plugging this back in, we get:

$$\begin{aligned}
\mathbb{E}_t\|x_{t+1} - x^*\|^2 &\leq \|x_t - x^*\|^2 + 2\tau_t[f(x^*) - f(x_t)] + \tau_t^2(\sigma^2 + 2\beta[f(x_t) - f(x^*)]) \\
\mathbb{E}\|x_{t+1} - x^*\|^2 &= \mathbb{E}[\mathbb{E}_t\|x_{t+1} - x^*\|^2] \qquad \text{tower property} \\
&\leq \mathbb{E}\|x_t - x^*\|^2 - 2\tau_t(1 - \beta\tau_t)\mathbb{E}[f(x_t) - f(x^*)] + \tau_t^2\sigma^2 \\
2\tau_t(1 - \beta\tau_t)\mathbb{E}[f(x_t) - f(x^*)] &\leq \mathbb{E}\|x_t - x^*\|^2 - \mathbb{E}\|x_{t+1} - x^*\|^2 + \tau_t^2\sigma^2
\end{aligned}$$

We can choose τ_t's such that for each t, $1 - \beta\tau_t \geq \frac{1}{2} \iff \tau_t \leq \frac{1}{2\beta}$. Assuming that, and summing the terms up,

$$\begin{aligned}
\sum_{t=1}^{T} \tau_t \mathbb{E}[f(x_t) - f(x^*)] &\leq \mathbb{E}\|x_0 - x^*\|^2 - \mathbb{E}\|x_{T+1} - x^*\|^2 + \sigma^2 \sum_{t=1}^{T} \tau_t^2 \\
&\leq \|x_0 - x^*\|^2 + \sigma^2 \sum_{t=1}^{T} \tau_t^2
\end{aligned}$$

Dividing by the sum of step sizes (assuming it to be finite):

$$\frac{\sum_{t=1}^{T} \tau_t \mathbb{E}[f(x_t) - f(x^*)]}{\sum_{t=1}^{T} \tau_t} = \mathbb{E}\left[\frac{\sum_{t=1}^{T} \tau_t f(x_t)}{\sum_{t=1}^{T} \tau_t}\right] - f(x^*) \leq \frac{\|x_0 - x^*\|^2}{\sum_{t=1}^{T} \tau_t} + \sigma^2 \frac{\sum_{t=1}^{T} \tau_t^2}{\sum_{t=1}^{T} \tau_t}$$

Define

$$\bar{x}_T = \frac{\sum_{t=1}^{T} \tau_t x_t}{\sum_{t=1}^{T} \tau_t}$$

. Thanks to convexity (Jensen's inequality) we have:

$$f(\bar{x}_T) \leq \frac{\sum_{t=1}^{T} \tau_t f(x_t)}{\sum_{t=1}^{T} \tau_t} \tag{2}$$

$$\mathbb{E}[f(\bar{x}_T) - f(x^*)] \leq \frac{\|x_0 - x^*\|^2}{\sum_{t=1}^{T} \tau_t} + \sigma^2 \frac{\sum_{t=1}^{T} \tau_t^2}{\sum_{t=1}^{T} \tau_t}$$

A set of sufficient conditions (Robbins-Monro) on the step size schedule in order for the gap to go to zero as $T \to \infty$ would be:

$$\sum_{t=1}^{\infty} \tau_t = \infty, \quad \sum_{t=1}^{\infty} \tau_t^2 < \infty$$

One example of a step size schedule for infinite time horizon that satisfies this is $\tau_t = \frac{1}{t}$, while note that a constant step size does not. In general for step size schedules of the form $\tau_t = t^{-a}$, we would want $\frac{1}{2} < a \leq 1$ to satisfy these conditions.

We can derive the optimal step size for constant horizon T. Let $\tau_t = \gamma(t)$ be the step size at time t and let $\Gamma = \sum_{t \leq T} \gamma(t)$. We can optimise the bound:

$$\frac{R^2 + \sigma^2 \sum_{t \leq T} \gamma^2(t)}{\Gamma}$$

This is minimised when $\sum_{t \leq T} \gamma^2(t)$ is minimised under the constraint $\sum_{t \leq T} \gamma(t) = \Gamma$. $T \sum_{t \leq T} \gamma^2(t) \geq (\sum \gamma(t))^2 = \Gamma^2$ by the Cauchy-Schwarz inequality with equality holding at constant $\gamma(t) = \frac{\Gamma}{T}$, which is thus the minimiser. Putting it back in:

$$\min_{\Gamma} \frac{R^2}{\Gamma} + \frac{\sigma^2 \Gamma}{T}$$

Minimising this we get $\Gamma = \frac{R\sqrt{T}}{\sigma}$ and thus $\tau_t = \min\{\frac{1}{2\beta}, \frac{R}{\sigma\sqrt{T}}\}$ (incorporating our eariler assumption). Note that the larger the variance of our gradient estimator, the smaller the step size we must take.

Thus, we have two regimes. For large T, specifically if $T \geq \left(\frac{2\beta R}{\sigma}\right)^2$, the rate of convergence we get is $O\left(\frac{R\sigma}{\sqrt{T}}\right)$- we get faster convergence when the variance is lower. Otherwise we clip $\tau_t = \frac{1}{2\beta}$ and get $\mathbb{E}[f(\bar{x}_T) - f(x^*)] \leq \frac{2\beta R^2}{T} + \frac{\sigma^2}{2\beta}$. When the noise is very low i.e. $\sigma \ll \beta$, (as in the extreme case when $\sigma = 0$) we are in the second regime and approach the $O(\frac{1}{T})$ rate of gradient descent.

For unknown T, we analyse $\frac{1}{2\beta\sqrt{t}}$ instead:

$$\mathbb{E}[f(\bar{x}_T) - f(x^*)] \leq \frac{\|x_0 - x^*\|^2}{\sum_{t=1}^{T} \frac{1}{2\beta\sqrt{t}}} + \sigma^2 \frac{\sum_{t=1}^{T} \frac{1}{4\beta^2 t}}{\sum_{t=1}^{T} \frac{1}{2\beta\sqrt{t}}}$$

Now, by the integral comparison test:

$$\int_1^T t^{-1/2} dt \leq \sum_{t=1}^{T} \frac{1}{\sqrt{t}} \leq 1 + \int_1^T t^{-1/2} dt$$

$$2(\sqrt{T} - 1) \leq \sum_{t=1}^{T} \frac{1}{\sqrt{t}} \leq 2\sqrt{T} + 1$$

$$\int_1^T t^{-1} dt \leq \sum_{t=1}^{T} \frac{1}{t} \leq 1 + \int_1^T t^{-1} dt$$

$$\log T \leq \sum_{t=1}^{T} \frac{1}{t} \leq 1 + \log T$$

$$\mathbb{E}[f(\bar{x}_T) - f(x^*)] \leq \frac{\beta\|x_0 - x^*\|^2}{\sqrt{T} - 1} + \frac{\sigma^2(1 + \log T)}{4\beta(\sqrt{T} - 1)}$$

□

If we have an almost perfect gradient estimator with small $\sigma^2 \ll \beta$ we get a $O(\frac{1}{\sqrt{T}})$ convergence rate. In other settings it is $O(\frac{\log T}{\sqrt{T}})$.

Corollary 3.2. *The SGD algorithm can be applied to the setting of parameter estimation for linear regression where the objective is of the form $f = \sum_{i=1}^{n} f_i$, and the gradient estimator is $\hat{\nabla} f = \nabla f_{i_t}, i_t \sim$ Unif$(\{1, \ldots, n\}$. It achieves convergence at the given rate assuming f is β-smooth and convex, boundedness of gradient estimator variance and the data points are IID.*

4 SGD for strongly convex smooth functions

Theorem 4.1. *If the assumptions for theorem 3.1 hold, and additionally we have f is α-strongly convex, then for the SGD update*

$$x_{t+1} := x_t - \frac{\hat{\nabla} f(x_t)}{\alpha t}$$

the expected function gap is

$$\mathbb{E}[f(\bar{x}_T) - f(x^*)] < \frac{4\sigma^2}{\alpha T} = O\left(\frac{1}{T}\right)$$

where

$$\tau_t = \frac{1}{\alpha t}, \bar{x}_T = \frac{\sum_{t=1}^{T} \tau_t x_t}{\sum_{t=1}^{T} \tau_t}$$

Recall that for a similar setting in the non-stochastic gradient descent case, we had a geometric $O\left(\left(1 - \frac{\alpha}{\beta}\right)^T\right)$ rate of convergence. In both cases, a more strongly convex (higher α) function yields faster convergence.

5 Stochastic Variance Reduced Gradient

In SGD, by using an unbiased estimator instead of the full gradient, we reduce the computational effort required per iteration but increase the number of iterations required by worsening the convergence rate. Note that the convergence rate linearly depends on the variance of our gradient estimator.

SVRG utilises a different unbiased estimator of the gradient. This estimator relies on a periodic "snapshots" $\tilde{x}$ i.e. iterates where the full gradient is evaluated and stored. The SVRG update is of the form:

$$x_{t+1} := x_t - \tau g_t$$

where

$$g_t := \hat{\nabla} f(x_t) + \nabla f(\tilde{x}) - \nabla f_{i_t}(\tilde{x})$$

The gradient estimator is unbiased:

$$\begin{aligned}
\mathbb{E}[\nabla f_{i_t}(x_t) + \nabla f(\tilde{x}) - \nabla f_{i_t}(\tilde{x})] &= \mathbb{E}[\nabla f_{i_t}(x_t)] + \nabla f(\tilde{x}) - \mathbb{E}[\nabla f_{i_t}(\tilde{x})] \\
&= \nabla f(x_t) + \nabla f(\tilde{x}) - \nabla f(\tilde{x}) \\
&= \nabla f(x_t)
\end{aligned}$$

The variance of g_t decreases as we converge towards the optimum (as opposed to that of SGD which stays constant):

Since $\tilde{x}$ is deterministic given all iterates upto time t, we have:

$$\begin{aligned}\text{Var}_t[g_t] &= \text{Var}_t[\nabla f_{i_t}(x_t) - \nabla f_{i_t}(\tilde{x})] \\ &\leq \mathbb{E}_t|\nabla f_{i_t}(x_t) - \nabla f_{i_t}(\tilde{x})|^2 \\ &\leq \beta^2|x_t - \tilde{x}|^2 \qquad \text{smoothness}\end{aligned}$$

So in the case of SVRG, we don't actually need a separate assumption about the boundedness of variance along with smoothness.

Hence in SVRG, we navigate the tradeoff of variance and gradient computation by periodically evaluating the full gradient and using it as a baseline to calibrate our gradient estimator. This reduces the variance as we converge towards the optimum and allows SVRG to achieve the same rates of convergence as full gradient descent (with different, worse constants).

Algorithm 1 SVRG

Require: Update frequency m and learning rate η
  Initialize $\tilde{x}_0$
  for $s = 1, 2, \ldots$ **do**
    $\tilde{x} = \tilde{x}_{s-1}$
    $\tilde{\mu} = \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(\tilde{x}) = \nabla f(\tilde{x})$
    $x_0 = \tilde{x}$
    for $t = 1, 2, \ldots, m$ **do**
      Sample uniformly $i_t \in \{1, \ldots, n\}$
      $x_t = x_{t-1} - \eta(\nabla f_{i_t}(x_{t-1}) - \nabla f_{i_t}(\tilde{x}) + \tilde{\mu})$
    end for
    Set $\tilde{x}_s = x_t$ for randomly chosen $t \in \{0, \ldots, m-1\}$
  end for

Theorem 5.1. *Let $f = \sum_{i \leq n} f_i$ be such that each f_i is β-smooth and convex, f is α-strongly convex. Then for algorithm 1, the expected function gap converges geometrically as*

$$\mathbb{E}[f(\tilde{x}_S) - f(x^*)] \leq k^S[f(\tilde{x}_0) - f(x^*)]$$

where

$$k = \frac{1}{\alpha\eta(1 - 2\beta\eta)m} + \frac{2\beta\eta}{1 - 2\beta\eta}$$

assuming m is large enough so that $k < 1$, and $S = \lfloor \frac{T}{m} \rfloor$ is the number of periods.

For the proof of this theorem and further reading about SVRG, refer to [2]. For smooth but not strongly convex problems, SVRG achieves a convergence rate of $O(\frac{1}{T})$. Unlike SGD, the learning rate of SVRG does not need to decay.

Note that k is a decreasing function of m, reflecting that we progress further in each epoch due to more steps. However, this is counteracted by the T/m exponent- the higher the value of m,

the fewer epochs we can complete for a fixed T. Also note that for $k(\eta) < 1$, it is necessary that $\eta < \frac{1}{4\beta}$.

We can find the optimal learning rate for a given α, β and m as follows:

$$\begin{aligned}
\frac{dk}{d\eta} = 0 &\implies -\frac{1-4\beta\eta}{\alpha m(\eta\phi)^2} + \frac{2\beta}{\phi^2} = 0, \quad \phi = 1 - 2\beta\eta \\
&\implies 2\alpha\beta m\,\eta^2 + 4\beta\,\eta - 1 = 0 \\
\implies \eta^* &= \frac{-4\beta + \sqrt{16\beta^2 + 8\alpha\beta m}}{4\alpha\beta m} \\
&= \frac{\sqrt{1 + \frac{\alpha}{2\beta}m} - 1}{\alpha\, m} \\
&\approx \frac{1}{\sqrt{2\alpha\beta m}} \quad \text{if } m \gg \frac{2\beta}{\alpha}
\end{aligned}$$

$k(\eta)$ is convex on the domain $(0, \frac{1}{2\beta})$, which covers all $\eta < \frac{1}{4\beta}$ satisfying the necessary condition. $\eta^* \in (0, \frac{1}{4\beta})$ so it is a local minimiser.

Now we want $k(\eta^*) < 1$ for convergence, so plugging back $\eta^* = \frac{\sqrt{1+\frac{\alpha}{2\beta}m}-1}{\alpha\, m}$

$$\begin{aligned}
k(\eta^*) &= \frac{1}{\alpha\eta(1-2\beta\eta)m} + \frac{2\beta\eta}{1-2\beta\eta} \\
&= \frac{1 + 2\alpha\beta m\eta^2}{\alpha m\eta(1-2\beta\eta)} \\
&= \frac{2 - 4\beta\eta^*}{\alpha m\eta^*(1-2\beta\eta^*)} \quad \text{(using } 2\alpha\beta m\eta^2 = 1 - 4\beta\eta \text{ at } \eta^*\text{)} \\
&= \frac{2}{\sqrt{1 + \frac{\alpha}{2\beta}m} - 1}
\end{aligned}$$

$$\begin{aligned}
k < 1 &\iff \frac{2}{\sqrt{1 + \frac{\alpha}{2\beta}m} - 1} < 1 \\
&\iff \sqrt{1 + \frac{\alpha}{2\beta}m} > 3 \\
&\iff \frac{\alpha}{2\beta}m > 8 \\
&\iff m > \frac{16\beta}{\alpha}
\end{aligned}$$

We can find an approximation (for $m/2\kappa \gg 1$ where $\kappa = \beta/\alpha$) for the optimal value of m by minimising $k^{T/m}$:

$$k = \frac{2}{\sqrt{1 + \frac{\alpha}{2\beta}m} - 1} \approx \frac{2}{\sqrt{\frac{m}{2\kappa}}} \approx \sqrt{\frac{8\kappa}{m}}$$
$$\log k^{T/m} = \frac{T}{2m}(\log 8\kappa - \log m)$$

Differentiating and setting derivative to zero we get $m^* = 8\kappa e$ and since it has to be an integer we take the best one out of $\lceil 8\kappa e \rceil, \lfloor 8\kappa e \rfloor$. The rate we get is $e^{-T/16\kappa e}$.

The lower bound on m is an artifact of the analysis suggesting that the the convergence bound is not tight; at $m = 1$ the algorithm reduces to gradient descent which we showed in class converged at a rate geometric $O((1 - \frac{\alpha}{\beta})^T)$. An alternative form of SVRG whose theoretical analysis does not require the period to scale with condition number has been analysed in [3].

A compilation of convergence results and derivations on different variants of SGD is available in [1].

References

[1] Guillaume Garrigos and Robert M. Gower. *Handbook of Convergence Theorems for (Stochastic) Gradient Methods*. 2024. arXiv: 2301.11235 [math.OC]. URL: https://arxiv.org/abs/2301.11235.

[2] Rie Johnson and Tong Zhang. "Accelerating stochastic gradient descent using predictive variance reduction". In: *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 1*. NIPS'13. Lake Tahoe, Nevada: Curran Associates Inc., 2013, pp. 315–323.

[3] Othmane Sebbouh et al. *Towards closing the gap between the theory and practice of SVRG*. 2021. arXiv: 1908.02725 [math.OC]. URL: https://arxiv.org/abs/1908.02725.