1 Proof of estimate sequence (continued)

So far we have proved that if at iteration k we have

- upper bound: $\phi_k(x) \leq (1-\lambda_k)\phi_k(x) + \lambda_k \phi_0(x)$
- lower bound: $f(x_k) \leq \phi_k(x)$

Then the same inequalities are true for iteration $k+1$ as well. To complete the induction we have to show that these inequalities are true for $k=0$. We said that at $k=0$ we have $\lambda_0 = 1$. Then the upper bound is satisfied since it is equivalent to

$$\phi_0(x) \leq (1-1)f(x) + \phi_0(x) = \phi_0(x),$$

which is true. The lower bound is satisfied because we defined $\phi_0(x) := f(x_0) + \frac{L}{2}\|x - x_0\|_2^2$, which implies that $\min_{x\in\mathbb{R}^n}\ \phi_0(x) = f(x_0)$. Thus the lower bound is equivalent to

$$f(x_0) \leq \min_{x\in\mathbb{R}^n}\ \phi_0(x) = f(x_0),$$

which is true.

2 Accelerated gradient (continued)

We proved that there exists an estimate sequence $\{\phi_k, \lambda_k, x_k\}_{k\in\mathbb{N}_0}$ with $\phi_0(x) := f(x_0) + \frac{L}{2}\|x - x_0\|_2^2$ and $\lambda_k \leq c/k^2$ for some constant $c > 0$. To prove this, we constructively build an estimate sequence based on Algorithm 1 (which is the algorithm that we call accelerated gradient).

```
Pick an arbitrary $x_0$
Set $\gamma_0 = 0$
Set $z_0 = x_0$
Set $\phi_0(x) := f(x_0) + \frac{L}{2}\|x - x_0\|_2^2$
while termination criterion is not satisfied do
    $y_{k-1} := (1-\gamma_k)x_{k-1} + \gamma_k z_{k-1}$
    $\phi_k(x) := (1-\gamma_k)\phi_{k-1} + \gamma_k(f(y_{k-1}) + \nabla f(y_{k-1})^T(x - y_{k-1}))$
    $z_k :=$ argmin$_{x\in\mathbb{R}^n}\ \phi_k(x)$
    $x_k := y_{k-1} - \frac{1}{L}\nabla f(y_{k-1})$
end
```

Algorithm 1: Accelerated gradient

2.1 How to pick γ_k

Note that you cannot pick γ_k arbitrarily. During the proof we restricted γ_k to satisfy

- $\gamma_k \in [0,1]$
- Let $\lambda_k := (1-\gamma_k)\lambda_{k-1}$ and $\gamma_0 = 0$. Then γ_k must satisfy $\lambda_k/\gamma_k^2 \geq 1$.
- $\lambda_k \leq c/k^2$ for some constant $c > 0$ (this is given in the preamble of the theorem for an estimate sequence).

One way of setting γ_k such that the above are satisfied is to set it as

- $\gamma_0 = \gamma_1 = \gamma_2 = \gamma_3 = 0$
- $\gamma_k = 2/k \, \forall k \geq 4.$

(Verify that this true by plugging in the definition of γ_k in the above constraints)
This selection of γ_k gives the accelerated gradient method in Algorithm 2

```
Pick an arbitrary $x_0$
Set $\gamma_0 = 0$
Set $z_0 = x_0$
Set $\phi_0(x) := f(x_0) + \frac{L}{2}\|x - x_0\|_2^2$
while termination criterion is not satisfied do
    $\gamma_k := \begin{cases} \frac{2}{k} & k \geq 4 \\ 0 & \text{otherwise} \end{cases}$
    $y_{k-1} := (1-\gamma_k)x_{k-1} + \gamma_k z_{k-1}$
    $\phi_k(x) := (1-\gamma_k)\phi_{k-1} + \gamma_k(f(y_{k-1}) + \nabla f(y_{k-1})^T(x - y_{k-1}))$
    $z_k := \text{argmin}_{x \in \mathbb{R}^n} \phi_k(x)$
    $x_k := y_{k-1} - \frac{1}{L}\nabla f(y_{k-1})$
end
```

Algorithm 2: Accelerated gradient (with γ_k defined explicitly)

2.2 Simplifying the step $z_k := \text{argmin}_{x \in \mathbb{R}^n} \phi_k(x)$

We will show in this subsection that the step

$$z_k := \text{argmin}_{x \in \mathbb{R}^n} \phi_k(x)$$

is equivalent to

$$z_k := z_{k-1} - \frac{\gamma_k}{\lambda_k}\frac{1}{L}\nabla f(y_{k-1}).$$

This will simplify Algorithm 2 even further into Algorithm 3.

```
Pick an arbitrary $x_0$
Set $\gamma_0 = 0$
Set $z_0 = x_0$
while termination criterion is not satisfied do
    $\gamma_k := \begin{cases} \frac{2}{k} & k \geq 4 \\ 0 & \text{otherwise} \end{cases}$
    $y_{k-1} := (1-\gamma_k)x_{k-1} + \gamma_k z_{k-1}$
    $z_k := z_{k-1} - \frac{\gamma_k}{\lambda_k}\frac{1}{L}\nabla f(y_{k-1})$
    $x_k := y_{k-1} - \frac{1}{L}\nabla f(y_{k-1})$
end
```

Algorithm 3: Accelerated gradient

Remember that in the proof of the estimate sequence we showed that

$$\phi_k(x) = \phi_k^* + \frac{L\lambda_k}{2}\|x - z_k\|_2^2 \, \forall k, \tag{1}$$

where ϕ_k^* is the minimum value of $\phi_k(x)$. From the definition $\phi_k(x)$ we also have that

$$\phi_k(x) := (1-\gamma_k)\phi_{k-1} + \gamma_k(f(y_{k-1}) + \nabla f(y_{k-1})^T(x - y_{k-1})). \tag{2}$$

Combining (1) and (2) we get

$$(1-\gamma_k)\phi_{k-1} + \gamma_k(f(y_{k-1}) + \nabla f(y_{k-1})^T(x - y_{k-1})) = \phi_k^* + \frac{L\lambda_k}{2}\|x - z_k\|_2^2. \tag{3}$$

Since (1) holds for any k we also have that

$$\phi_{k-1}(x) = \phi_{k-1}^* + \frac{L\lambda_k}{2}\|x - z_{k-1}\|_2^2$$

Using this in (3) we get that the LHS in (3) is equivalent to

$$(1-\gamma_k)\left(\phi_{k-1}^* + \frac{L\lambda_{k-1}}{2}\|x - z_{k-1}\|_2^2\right) + \gamma_k(f(y_{k-1}) + \nabla f(y_{k-1})^T(x - y_{k-1}))$$

Putting everything together we have that

$$(1-\gamma_k)\left(\phi_{k-1}^* + \frac{L\lambda_{k-1}}{2}\|x - z_{k-1}\|_2^2\right) + \gamma_k(f(y_{k-1}) + \nabla f(y_{k-1})^T(x - y_{k-1})) = \\ \phi_k^* + \frac{L\lambda_k}{2}\|x - z_k\|_2^2$$

Taking derivatives in both sides we get

$$(1-\gamma_k)\lambda_{k-1}L(x - z_{k-1}) + \gamma_k\nabla f(y_{k-1}) = \lambda_k L(x - z_k),$$

which is equivalent to

$$\lambda_k L(x - z_{k-1}) + \gamma_k \nabla f(y_{k-1}) = \lambda_k L(x - z_k).$$

Dividing by $\lambda_k L$ we get

$$x - z_{k-1} + \frac{\gamma_k}{\lambda_k L}\nabla f(y_{k-1}) = x - z_k.$$

Removing x and solving w.r.t z_k we get

$$z_k = z_{k-1} - \frac{\gamma_k}{\lambda_k L}\nabla f(y_{k-1}).$$

3 Assumptions for convergence rate of stochastic gradient

- f is bounded below (not necessarily convex)
- f_i is differentiable $\forall i = 1, \ldots, n$
- $\nabla f(x) = \frac{1}{n}\sum_{i=1}^n \nabla f_i(x)$ is Lipschitz continuous with constant L
- $\mathbb{E}\left[\|\nabla f_i(x_k)\|_2^2 \mid x_k\right] \leq B^2$

4 Proof for convergence rate of stochastic gradient

From the Fundamental Theorem of Calculus (FToC) we have that

$$f(x_{k+1}) \le f(x_k) + \nabla f(x_k)^T (x_{k+1} - x_k) + \frac{L}{2}\|x_{k+1} - x_k\|_2^2$$

Using the definition of stochastic gradient

$$x_{k+1} = x_k - \alpha_k \nabla f_i(x_k)$$

We get

$$f(x_{k+1}) \le f(x_k) - \alpha_k \nabla f(x_k)^T \nabla f_i(x_k) + \frac{\alpha_k^2 L}{2}\|\nabla f_i(x_k)\|_2^2$$

Note that x_{k+1} and $\nabla f_i(x_k)$ are random variables since we choose sample i randomly at each iteration of stochastic gradient. Let's assume that we pick sample uniformly at random

$$p(i_j = i) = \frac{1}{n} \ \forall k.$$

Let's use conditional expectation w.r.t x_k and the above uniform distribution, then we get

$$\begin{aligned} \mathbb{E}[f(x_{k+1}) \mid x_k] &\le \mathbb{E}\left[f(x_k) - \alpha_k \nabla f(x_k)^T \nabla f_i(x_k) + \frac{\alpha_k^2 L}{2}\|\nabla f_i(x_k)\|_2^2 \mid x_k\right] \\ &= f(x_k) - \alpha_k \nabla f(x_k)^T \mathbb{E}\left[\nabla f_i(x_k) \mid x_k\right] + \frac{\alpha_k^2 L}{2}\mathbb{E}\left[\|\nabla f_i(x_k)\|_2^2 \mid x_k\right] \end{aligned} \tag{4}$$

The second equality is due to linearity of expectation and the fact that we assume that α_k is not a random variable. Furthermore, we have

$$\mathbb{E}\left[\nabla f_i(x_k) \mid x_k\right] = \sum_{i=1}^n p(i_k = i)\nabla f_i(x_k) = \sum_{i=1}^n \frac{1}{n}\nabla f_i(x_k) = \frac{1}{n}\sum_{i=1}^n \nabla f_i(x_k) = \nabla f(x_k). \tag{5}$$

This means that if we use uniform probabilities at each iteration k then $\nabla f_i(x_k)$ is an unbiased estimator of $\nabla f(x_k)$. Combining (4) and (5) we get

$$\mathbb{E}[f(x_{k+1}) \mid x_k] \le f(x_k) - \alpha_k \|\nabla f(x_k)\|_2^2 + \frac{\alpha_k^2 L}{2}\mathbb{E}\left[\|\nabla f_i(x_k)\|_2^2 \mid x_k\right] \tag{6}$$

Using the forth assumption we get that

$$\mathbb{E}[f(x_{k+1}) \mid x_k] \le f(x_k) - \alpha_k \|\nabla f(x_k)\|_2^2 + \frac{\alpha_k^2 L}{2}B^2. \tag{7}$$

Let's re-arrange to get

$$\alpha_k \|\nabla f(x_k)\|_2^2 \le f(x_k) - \mathbb{E}[f(x_{k+1}) \mid x_k] + \frac{\alpha_k^2 L}{2}B^2.$$

Take expectation over the first $t+1$ iterations

$$\mathbb{E}\left[\alpha_k \|\nabla f(x_k)\|_2^2\right] \le \mathbb{E}\left[f(x_k) - \mathbb{E}[f(x_{k+1}) \mid x_k] + \frac{\alpha_k^2 L}{2}B^2\right].$$

Note that this expectation is over whatever has happened up to iteration $t+1$ for some integer t. Using the assumption that α_k is not a random variable and linearity of expectation we get

$$\alpha_k \mathbb{E}\left[\|\nabla f(x_k)\|_2^2\right] \leq \mathbb{E}\left[f(x_k)\right] - \mathbb{E}\left[f(x_{k+1})\right] + \frac{\alpha_k^2 L}{2} B^2.$$

Let's sum the above inequality over the first t iterations, we get

$$\sum_{i=0}^{t} \alpha_i \mathbb{E}\left[\|\nabla f(x_i)\|_2^2\right] \leq \sum_{i=0}^{t} \left(\mathbb{E}\left[f(x_i)\right] - \mathbb{E}\left[f(x_{i+1})\right]\right) + \frac{LB^2}{2} \sum_{i=0}^{t} \alpha_i^2.$$

The sum in the RHS telescopes, which gives

$$\sum_{j=0}^{t} \alpha_j \mathbb{E}\left[\|\nabla f(x_j)\|_2^2\right] \leq \mathbb{E}\left[f(x_0)\right] - \mathbb{E}\left[f(x_{t+1})\right] + \frac{LB^2}{2} \sum_{j=0}^{t} \alpha_j^2.$$

Using $\mathbb{E}\left[f(x_{t+1})\right] \geq f^*$ and the fact that $f(x_0)$ is not a random variable we get

$$\sum_{j=0}^{t} \alpha_j \mathbb{E}\left[\|\nabla f(x_j)\|_2^2\right] \leq f(x_0) - f^* + \frac{LB^2}{2} \sum_{j=0}^{t} \alpha_j^2.$$

Further lower bounding the LHS we get

$$\min_{j=1,\ldots,t} \mathbb{E}\left[\|\nabla f(x_j)\|_2^2\right] \sum_{j=0}^{t} \alpha_j \leq f(x_0) - f^* + \frac{LB^2}{2} \sum_{j=0}^{t} \alpha_j^2$$

Let's divide by $\sum_{j=0}^{t} \alpha_j$ to get

$$\min_{j=1,\ldots,t} \mathbb{E}\left[\|\nabla f(x_j)\|_2^2\right] \leq \frac{f(x_0) - f^*}{\sum_{j=0}^{t} \alpha_j} + \frac{LB^2}{2} \frac{\sum_{j=0}^{t} \alpha_j^2}{\sum_{j=0}^{t} \alpha_j}.$$

We now have to make some decisions about selecting α_j's.

- If $B^2 = 0$ then by setting $\alpha_j = 1/L$ we get the usual rate of $\mathcal{O}(1/t)$ for gradient descent.
- Decreasing step-sizes: if $\alpha_j = 1/i$ then $\sum_{j=0}^{t} \alpha_j = \mathcal{O}(\log t)$ and $\sum_{j=0}^{t} \alpha_j^2 = \mathcal{O}(1)$, which gives rate $\mathcal{O}(1/\log t)$.
- Large decreasing step-size: if $\alpha_j = 1/\sqrt{j}$ then $\sum_{j=0}^{t} \alpha_j = \mathcal{O}(\sqrt{t})$ and $\sum_{j=0}^{t} \alpha_j^2 = \mathcal{O}(\log t)$, which gives rate $\mathcal{O}(\log t/\sqrt{t})$.
- Constant step-size: if $\alpha_j = \alpha$ for some constant α, then $\sum_{j=0}^{t} \alpha_j = (t+1)\alpha$ and $\sum_{j=0}^{t} \alpha_j^2 = (t+1)\alpha^2$, which gives rate $\mathcal{O}(1/t + \alpha)$. This means implies that the minimum expected norm of the gradient will never go to zero and then algorithm only converges to a neighborhood of a stationary point.