NOTES ON FIRST-ORDER METHODS FOR MINIMIZING SMOOTH FUNCTIONS

1. Introduction. We consider first-order methods for smooth, unconstrained optimization:

$$\underset{x\in\mathbf{R}^n}{\text{minimize}}\, f(x), \tag{1.1}$$

where $f : \mathbf{R}^n \to \mathbf{R}$. We assume the optimum of f exists and is unique. By first-order, we are referring to methods that use only function value and gradient information. Compared to second-order methods (e.g. Newton-type methods, interior-point methods), first-order methods usually converge slower, and their performance degrades on ill-conditioned problems. However, they are well-suited to large-scale problems because each iteration usually only requires vector operations,[1] in addition to the evaluation of $f(x)$ and $\nabla f(x)$.

2. Gradient descent. The simplest first-order method is *gradient descent* or *steepest descent*. It starts at an initial point x_0 and generates successive iterates by

$$x_{k+1} \leftarrow x_k - \alpha_k \nabla f(x_k). \tag{2.1}$$

The step size α_k is either fixed or chosen by a line search that (approximately) minimizes $f(x_k - \alpha\nabla f(x_k))$. For small to medium-scale problems, each iteration is inexpensive, requiring only vector operations. However, the iterates may converge slowly to the optimum x^*.

Let's take a closer look at the convergence of gradient descent. For now, fix a step size α and consider gradient descent as a fixed point iteration:

$$x_{k+1} \leftarrow G_\alpha(x_k),$$

where $G_\alpha(x) := x - \alpha\nabla f(x)$ is a contraction:

$$\|G_\alpha(x) - G_\alpha(y)\|_2 \leq L_G\|x - y\|_2 \text{ for some } L_G < 1.$$

Since the optimum x^* is a fixed point of G_α, we expect the fixed point iteration $x_{k+1} \leftarrow G_\alpha(x_k)$ to converge to the fixed point.

[1] Second-order methods usually require the solution of linear system(s) at each iteration, making them prohibitively expensive for very large problems.

LEMMA 2.1. *When the gradient step $G_\alpha(x)$ is a contraction, gradient descent converges linearly to x^* :*

$$\|x_{k+1} - x^*\|_2 \leq L_G^k \|x_0 - x^*\|_2.$$

PROOF. We begin by showing that each iteration of gradient descent is a contraction:

$$\begin{aligned} \|x_{k+1} - x^*\|_2 &= \|x_k - \alpha_k \nabla f(x_k) - (x^* - \alpha_k \nabla f(x^*))\|_2 \\ &= \|G_\alpha(x) - G_\alpha(x^*)\|_2 \\ &\leq L_G \|x_k - x^*\|_2. \end{aligned}$$

By induction, $\|x_{k+1} - x^*\|_2 \leq L_G^k \|x_0 - x^*\|_2$. □

At first blush, it seems like gradient descent converges quickly. However, the contraction coefficient L_G depends poorly on the condition number $\kappa := \frac{L}{\mu}$ of the problem. In particular $L_G \sim 1 - \frac{1}{\kappa}$ with an optimal step size. From now on, we assume the function f is convex.

LEMMA 2.2. *When $f : \mathbf{R}^n \to \mathbf{R}$ is*

1. *twice continuously differentiable*
2. *strongly convex with constant $\mu > 0$:*

$$f(y) \geq f(x) + \nabla f(x)^T (y - x) + \tfrac{\mu}{2} \|y - x\|_2^2 \tag{2.2}$$

3. *strongly smooth with constant $L \geq 0$:*

$$f(y) \leq f(x) + \nabla f(x)^T (y - x) + \frac{L}{2} \|y - x\|_2^2, \tag{2.3}$$

the contraction coefficient L_G at most $\max\{|1 - \alpha\mu|, |1 - \alpha L|\}$.

PROOF. By the mean value theorem,

$$\begin{aligned} \|G_\alpha(x) - G_\alpha(y)\|_2 &= \|x - \alpha \nabla f(x) - (y - \alpha \nabla f(y))\|_2 \\ &= \|(I - \nabla^2 f(z))(x - y)\|_2 \\ &\leq \|I - \nabla^2 f(z)\|_2 \|x - y\|_2 \end{aligned}$$

for some z on the line segment between x and y. By strong convexity and smoothness, the eigenvalues of $\nabla^2 f(z)$ are between μ and L. Thus,

$$\|I - \alpha \nabla^2 f(z)\|_2 \leq \max\{|1 - \alpha\mu|, |1 - \alpha L|\}.$$

□

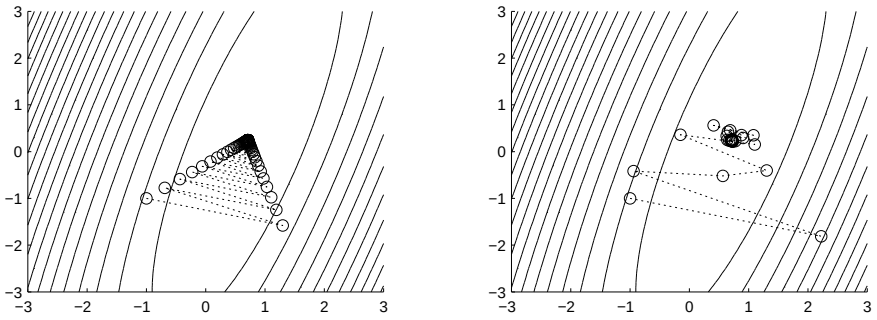

FIG 1. *The iterates of gradient descent (left panel) and the heavy ball method (right panel) starting at* $(-1, -1)$.

We combine Lemmas 2.1 and 2.2 to deduce

$$\|x_{k+1} - x^*\|_2 \leq \max\{|1 - \alpha\mu|, |1 - \alpha L|\}^k \|x_0 - x^*\|_2.$$

The step size $\frac{2}{L+\mu}$ minimizes the contraction coefficient. Substituting the optimal step size, we conclude that

$$\|x_{k+1} - x^*\|_2 \leq \left(\frac{\kappa - 1}{\kappa + 1}\right)^k \|x_0 - x^*\|_2,$$

where $\kappa := \frac{L}{\mu}$ is a condition number of the problem. When κ is large, the contraction coefficient is roughly $1 - \frac{1}{\kappa}$. In other words, gradient descent requires at most $O\left(\kappa \log(\frac{1}{\epsilon})\right)$ iterations to attain an ϵ-accurate solution.

3. The heavy ball method. The heavy ball method is usually attributed to Polyak (1964). The intuition is simple: the iterates of gradient descent tend to bounce between the walls of narrow "valleys" on the objective surface. The left panel of Figure 2 shows the iterates of gradient descent bouncing from wall to wall. To avoid bouncing, the heavy ball method adds a momentum term to the gradient step:

$$x_{k+1} \leftarrow x_k - \alpha_k \nabla f(x_k) + \beta_k (x_k - x_{k-1}).$$

The term $x_k - x_{k-1}$ nudges x_{k+1} in the direction of the previous step (hence momentum). The right panel of Figure 2 shows the effects of adding momentum.

Let's study the heavy ball method and compare it to gradient descent. Since each iterate depends on the previous two iterates, we study the three-term recurrence in matrix form:

$$\begin{aligned}\left\|\begin{bmatrix} x_{k+1}-x^* \\ x_k - x^*\end{bmatrix}\right\|_2 &= \left\|\begin{bmatrix} x_k + \beta(x_k - x_{k-1}) - x^* \\ x_k - x^*\end{bmatrix} - \alpha\begin{bmatrix}\nabla f(x_k)\\ 0\end{bmatrix}\right\|_2 \\ &= \left\|\begin{bmatrix}(1+\beta)I & -\beta I \\ I & 0\end{bmatrix}\begin{bmatrix} x_k - x^* \\ x_{k-1}-x^*\end{bmatrix} - \alpha\begin{bmatrix}\nabla^2 f(z_k)(x_k - x^*)\\ 0\end{bmatrix}\right\|_2\end{aligned}$$

for some z_k on the line segment between x_k and x^*

$$\begin{aligned}&= \left\|\begin{bmatrix}(1+\beta)I - \alpha\nabla^2 f(z_k) & -\beta I \\ I & 0\end{bmatrix}\begin{bmatrix} x_k - x^* \\ x_{k-1}-x^*\end{bmatrix}\right\|_2 \\ &\le \left\|\begin{bmatrix}(1+\beta)I - \alpha_k\nabla^2 f(z_k) & -\beta I \\ I & 0\end{bmatrix}\right\|_2 \left\|\begin{bmatrix} x_k - x^* \\ x_{k-1}-x^*\end{bmatrix}\right\|_2. \qquad (3.1)\end{aligned}$$

The contraction coefficient is given by $\left\|\begin{bmatrix}(1+\beta)I - \alpha_k\nabla^2 f(z_k) & -\beta I \\ I & 0\end{bmatrix}\right\|$. We introduce a lemma to bound the norm.

LEMMA 3.1. *Under the conditions of Lemma 2.2,*

$$\left\|\begin{bmatrix}(1+\beta)I - \alpha\nabla^2 f(z) & -\beta I \\ I & 0\end{bmatrix}\right\|_2 \le \max\{|1-\sqrt{\alpha\mu}|, |1-\sqrt{\alpha L}|\}$$

for $\beta = \max\{|1-\sqrt{\alpha\mu}|, |1-\sqrt{\alpha L}|\}^2$.

PROOF. Let $\Lambda := \mathbf{diag}(\lambda_1, \ldots, \lambda_n)$, where $\{\lambda_i\}_{i\in[1:n]}$ are the eigenvalues of $\nabla^2 f(z)$. By diagonalization,

$$\begin{aligned}&\left\|\begin{bmatrix}(1+\beta)I - \alpha\nabla^2 f(z) & -\beta I \\ I & 0\end{bmatrix}\right\|_2 \\ &\quad= \left\|\begin{bmatrix}(1+\beta)I - \alpha\Lambda & -\beta I \\ I & 0\end{bmatrix}\right\|_2 = \max_{i\in[1:n]}\left\|\begin{bmatrix}1+\beta-\alpha\lambda_i & -\beta \\ 1 & 0\end{bmatrix}\right\|_2.\end{aligned}$$

The second equality holds because it is possible to permute the matrix to a block diagonal matrix with 2×2 blocks. For each $i \in [1:n]$, the eigenvalues of the 2×2 matrices are given by the roots of

$$p_i(\lambda) := \lambda^2 - (1+\beta-\alpha\lambda_i)\lambda + \beta = 0.$$

It is possible to show that when β is at least $(1-\sqrt{\alpha\lambda_i})^2$, the roots of p_i are imaginary and both have magnitude β. Since $\beta := \max\{|1-\sqrt{\alpha\mu}|, |1-\sqrt{\alpha L}|\}^2$, the magnitude of the roots are at most β. □

THEOREM 3.2. *Under the conditions of Lemma 2.2, the heavy ball method with fixed parameters* $\alpha = \frac{4}{(\sqrt{L}+\sqrt{\mu})^2}$ *and* $\beta = \max\{|1-\sqrt{\alpha\mu}|, |1-\sqrt{\alpha L}|\}^2$ *converges linearly to* x^* :

$$\|x_{k+1} - x^*\|_2 \leq \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^k \|x_0 - x^*\|_2,$$

where $\kappa = \frac{L}{\mu}$.

PROOF. By Lemma 3.1,

$$\left\| \begin{bmatrix} x_{k+1} - x^* \\ x_k - x^* \end{bmatrix} \right\|_2 \leq \max\{|1-\sqrt{\alpha\mu}|, |1-\sqrt{\alpha L}|\} \left\| \begin{bmatrix} x_k - x^* \\ x_{k-1} - x^* \end{bmatrix} \right\|_2$$

Substituting $\alpha = \frac{4}{(\sqrt{L}+\sqrt{\mu})^2}$ gives

$$\left\| \begin{bmatrix} x_{k+1} - x^* \\ x_k - x^* \end{bmatrix} \right\|_2 \leq \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1} \left\| \begin{bmatrix} x_k - x^* \\ x_{k-1} - x^* \end{bmatrix} \right\|_2,$$

where $\kappa := \frac{L}{\mu}$. We induct on k to obtain the stated result. □

The heavy ball method also converges linearly, but its contraction coefficient is roughly $1 - \frac{1}{\sqrt{\kappa}}$. In other words, the heavy ball method attains an ϵ-accurate solution after at most $O\left(\sqrt{\kappa}\log(\frac{1}{\epsilon})\right)$ iterations. Figure 3 compares the convergence of gradient descent and the heavy ball method. Unlike gradient descent, the heavy ball method is *not* a descent method, i.e. $f(x_{k+1})$ is not necessarily smaller than $f(x_k)$.

On a related note, the conjugate gradient method (CG) is an instance of the heavy ball method with adaptive step sizes and momentum parameters. However, CG does not require knowledge of L and μ to choose step sizes.

4. Accelerated gradient methods.

4.1. *Gradient descent in the absence of strong convexity.* Before delving into Nesterov's celebrated method, we study gradient descent in the setting he considered. Nesterov studied (1.1) when f is continuously differentiable and strongly smooth. In the absence of strong convexity, (1.1) may not have a unique optimum, so we cannot expect the iterates to always converge to a point x^*.[2] However, if f remains convex, we expect the optimality gap $f(x_k) - f^*$ to converge to zero. To begin, we study the convergence rate of gradient descent in the absence of strong convexity.

[2]The iterates may converge to different points depending on the initial point x_0.

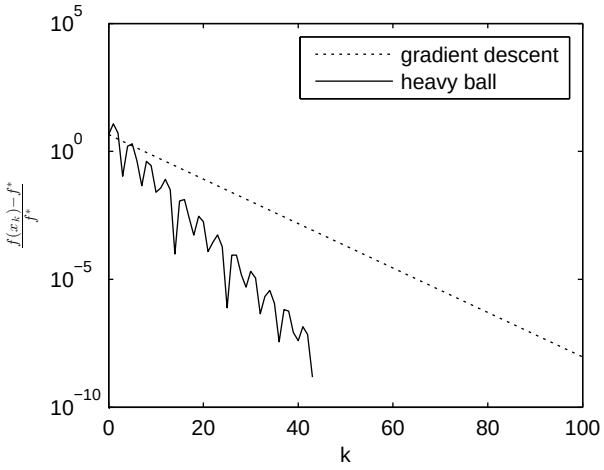

FIG 2. *Convergence of gradient descent and heavy ball function values on a strongly convex function. Although the heavy ball function values are not monotone, they converge faster.*

THEOREM 4.1. *When $f : \mathbf{R}^n \to \mathbf{R}$ is convex and*

1. *continuously differentiable*
2. *strongly smooth with constant L,*

gradient descent with step size $\frac{1}{L}$ satisfies

$$f(x_k) - f^* \leq \frac{L}{2k}\|x_0 - x^*\|_2^2.$$

PROOF. The proof consists of two parts. First, we show that, at each step, gradient descent reduces the function value by at least a certain amount. In other words, we show each iteration of gradient descent achieves *sufficient descent*. Recall $x_{k+1} \leftarrow x_k - \alpha\nabla f(x_k)$. By the strong smoothness of f,

$$f(x_{k+1}) \leq f(x_k) + \left(\frac{\alpha^2 L}{2} - \alpha\right)\|\nabla f(x_k)\|_2^2.$$

When $\alpha = \frac{1}{L}$, the right side simplifies to

$$f(x_{k+1}) \leq f(x_k) - \frac{1}{2L}\|\nabla f(x_k)\|_2^2. \tag{4.1}$$

The second part of the part of the proof combines the sufficient descent condition with the convexity of f to obtain a bound on the optimality gap.

By the first-order condition for convexity,

$$\begin{aligned} f(x_{k+1}) &\leq f^* + \nabla f(x_k)^T(x_k - x^*) - \frac{1}{2L}\|\nabla f(x_k)\|_2^2 \\ &= f^* - \left(\frac{L}{2}\|x_k - x^*\|_2^2 - \nabla f(x_k)^T(x_k - x^*) + \frac{1}{2L}\|\nabla f(x_k)\|_2^2\right) + \frac{L}{2}\|x_k - x^*\|_2^2, \end{aligned}$$

where the second equality follows from completing the square,

$$\begin{aligned} &= f^* - \frac{L}{2}\|x_k - x^* - \frac{1}{L}\nabla f(x_k)\|_2^2 + \frac{L}{2}\|x_k - x^*\|_2^2 \\ &= f^* - \frac{L}{2}\|x_{k+1} - x^*\|_2^2 + \frac{L}{2}\|x_k - x^*\|_2^2. \end{aligned}$$

We sum over iterations to obtain

$$\begin{aligned} \sum_{l=1}^{k} f(x_l) - f^* &\leq \frac{L}{2}\sum_{l=1}^{k} \|x_{l-1} - x^*\|_2^2 - \|x_l - x^*\|_2^2 \\ &= \frac{L}{2}\left(\|x_0 - x^*\|_2^2 - \|x_k - x^*\|_2^2\right) \\ &\leq \frac{L}{2}\|x_0 - x^*\|_2^2. \end{aligned}$$

Since $f(x_k)$ is non-increasing, we have

$$f(x_k) - f^* \leq \frac{1}{k}\sum_{l=1}^{k} f(x_l) - f^* \leq \frac{L}{2k}\|x_0 - x^*\|_2^2.$$

□

In the absence of strong convexity, the convergence rate of gradient descent slows to sublinear. Roughly speaking, gradient descent requires at most $O(\frac{L}{\epsilon})$ iterations to obtain an ϵ-optimal solution.

The story is similar when the step sizes are not fixed, but chosen by a backtracking line search. More precisely, a backtracking line search

1. starts with some initial trial step size $\alpha_k \leftarrow \alpha_{\text{init}}$
2. decreases the trial step size geometrically: $\alpha_k \leftarrow \frac{\alpha_k}{2}$[3]
3. continues until α_k satisfies a sufficient descent condition:

$$f(x_k + \alpha_k \nabla f(x_k)) \leq f(x_k) - \frac{\alpha_k}{2}\|\nabla f(x_k)\|_2^2. \tag{4.2}$$

[3]The new trial step may be $\beta \cdot \alpha$, for any $\beta \in (0, 1)$. In practice, it's frequently set close to 1 to avoid taking conservative step sizes. We let $\beta = \frac{1}{2}$ for simplicity.

LEMMA 4.2. *Under the conditions of Theorem 4.1, backtracking with initial step size α_{init} chooses a step size that is at least* $\min\{\alpha_{init}, \frac{1}{2L}\}$.

PROOF. Recall the quadratic upper bound

$$f(x_k + \alpha \nabla f(x_k)) \leq f(x_k) + \left(\frac{\alpha^2 L}{2} - \alpha\right) \|\nabla f(x_k)\|_2^2 \text{ for any } \alpha \geq 0.$$

It's minimized at $\alpha = \frac{1}{L}$. We check that choosing any $\alpha_k \leq \frac{1}{L}$ satisfies (4.2):

$$\begin{aligned} f(x_k + \alpha_k \nabla f(x_k)) &\leq f(x_k) + \frac{1}{L}\left(\frac{\alpha_k L}{2} - 1\right) \|\nabla f(x_k)\|_2^2 \\ &\leq f(x_k) + \frac{1}{L}\left(\frac{1}{2} - 1\right) \|\nabla f(x_k)\|_2^2 \\ &\leq f(x_k) - \frac{1}{2L}\|\nabla f(x_k)\|_2^2 \\ &\leq f(x_k) - \frac{\alpha_k}{2}\|\nabla f(x_k)\|_2^2. \end{aligned}$$

Since any step size shorter than $\frac{1}{L}$ satisfies (4.2), the line search stops backtracking at a step size longer than $\frac{1}{2L}$. □

Thus, at each step, gradient descent with backtracking line search decreases the function value by at least $\frac{1}{4L}\|\nabla f(x_k)\|_2^2$. The rest of the story is similar to that of gradient descent with a fixed step size. To summarize, gradient descent with a backtracking line search also requires at most $O(\frac{L}{\epsilon})$ iterations to obtain an ϵ-optimal solution.

4.2. *Nesterov's accelerated gradient method.* Accelerated gradient methods obtain an ϵ-optimal solution in $O(\sqrt{L/\epsilon})$ iterations. Nesterov (1983) is the original and (one of) the best known. It initializes $y_1 \leftarrow x_0, \gamma_0 \leftarrow 1$ and generates iterates according to

$$y_{k+1} \leftarrow x_k + \beta_k(x_k - x_{k-1}) \tag{4.3}$$

$$x_{k+1} \leftarrow y_{k+1} - \frac{1}{L}\nabla f(y_{k+1}), \tag{4.4}$$

where L is the strong smoothness constant of f and

$$\gamma_k \leftarrow \frac{1}{2}\left(\left(4\gamma_{k-1}^2 + \gamma_{k-1}^4\right)^{\frac{1}{2}} - \gamma_{k-1}^2\right) \tag{4.5}$$

$$\beta_k \leftarrow \gamma_k\left(1 - \gamma_{k-1}^{-1}\right). \tag{4.6}$$

It's very similar to the heavy ball method, except the gradient is evaluated at an intermediate point y_{k+1}. Often, the γ-update (4.5) is stated as the solution to

$$\gamma_k^2 = (1-\gamma_k)\gamma_{k-1}^2. \tag{4.7}$$

To study Nesterov's 1983 method, we first formulate the recursion in terms of interpretable quantities

$$z_k \leftarrow x_{k-1} + \frac{1}{\gamma_{k-1}}(x_k - x_{k-1}) : z_0 \leftarrow x_0 \tag{4.8}$$

$$y_{k+1} \leftarrow (1-\gamma_k)x_k + \gamma_k z_k \tag{4.9}$$

$$x_{k+1} \leftarrow y_{k+1} - \frac{1}{L}\nabla f(y_{k+1}). \tag{4.10}$$

Behind the scenes, the method forms an *estimating sequence* of quadratic functions centered at $\{z_k\}$:

$$\phi_k(x) := \phi_k^* + \frac{\eta_k}{2}\|x - z_k\|_2^2.$$

DEFINITION 4.3. *A sequence of functions $\phi_k : \mathbf{R}^n \to \mathbf{R}$ is an estimating sequence of $f : \mathbf{R}^n \to \mathbf{R}$ if, for some non-negative sequence of $\tilde{\eta}_k \to 0$,*

$$\phi_k(x) \leq (1-\tilde{\eta}_k)f(x) + \tilde{\eta}_k\phi_0(x).$$

Intuitively, an estimating sequence for f consists of successively tighter approximations of f. As we shall see, $\{\phi_k\}$ is an estimating sequence of f. Nesterov chose ϕ_0 to be a quadratic function centered at the initial point:

$$\phi_0(x) \leftarrow \phi_0^* + \frac{\eta_0}{2}\|x - x_0\|_2^2,$$

where $\phi_0^* = f(x_0)$ and $\eta_0 = L$. The subsequent ϕ_k are convex combinations of ϕ_{k-1} and the first-order approximation of f at y_{k+1} :

$$\phi_{k+1}(x) \leftarrow (1-\gamma_k)\phi_k(x) + \gamma_k\hat{f}_k(x), \tag{4.11}$$

where $\hat{f}_k(x) := f(y_{k+1}) + \nabla f(y_{k+1})^T(x - y_{k+1})$ and x_k, y_k are given by (4.10) and (4.9). It's straightforward to show the subsequent ϕ_k are quadratic.

LEMMA 4.4. *The functions ϕ_k have the form $\phi_k^* + \frac{\eta_k}{2}\|x - z_k\|_2^2$, where*

$$\eta_{k+1} \leftarrow (1-\gamma_k)\eta_k, \tag{4.12}$$

$$z_{k+1} \leftarrow z_k - \frac{\gamma_k}{\eta_{k+1}}\nabla f(y_{k+1}), \tag{4.13}$$

$$\phi_{k+1}^* \leftarrow (1-\gamma_k)\phi_k^* + \gamma_k\hat{f}_k(z_k) - \frac{\gamma_k^2}{2\eta_{k+1}}\|\nabla f(y_{k+1})\|_2^2. \tag{4.14}$$

PROOF. We know $\nabla^2\phi_k = \eta_k$, but by (4.11), it is also

$$(1-\gamma_k)\nabla^2\phi_{k-1} + \gamma_k\nabla^2\hat{f}_{k-1} = (1-\gamma_k)\eta_{k-1}I.$$

Thus $\eta_k \leftarrow (1-\gamma_k)\eta_{k-1}$. Similarly, we know $\nabla\phi_k = \eta_k(x - z_k)$, but it is also

$$\begin{aligned}
&(1-\gamma_k)\nabla\phi_{k-1} + \gamma_k\nabla\hat{f}_{k-1} \\
&\quad = (1-\gamma_k)\eta_{k-1}(x - z_{k-1}) + \gamma_k\nabla f(y_k) \\
&\quad = \eta_k(x - z_k) + \gamma_k\nabla f(y_k) \\
&\quad = \eta_k\left(x - \left(z_k - \frac{\gamma_k}{\eta_k}\nabla f(y_k)\right)\right).
\end{aligned}$$

Thus $z_k \leftarrow z_{k-1} - \frac{\gamma_{k-1}}{\eta_k}\nabla f(y_k)$. The intercept ϕ_k^* is simply $\phi_k(z_k)$:

$$\begin{aligned}
\phi_k(z_k) &= (1-\gamma_{k-1})\phi_{k-1}(z_k) + \gamma_{k-1}\hat{f}_{k-1}(z_k) \\
&= (1-\gamma_{k-1})\left(\phi_{k-1}^* + \frac{\eta_{k-1}}{2}\|z_k - z_{k-1}\|_2^2\right) + \gamma_{k-1}\hat{f}_{k-1}(z_k).
\end{aligned}$$

Since $z_k \leftarrow z_{k-1} - \frac{\gamma_{k-1}}{\eta_k}\nabla f(y_k)$, $\eta_k \leftarrow (1-\gamma_k)\eta_{k-1}$,

$$\begin{aligned}
&= (1-\gamma_{k-1})\phi_{k-1}^* + \frac{(1-\gamma_{k-1})\eta_{k-1}}{2}\left\|\frac{\gamma_{k-1}}{\eta_k}\nabla f(y_k)\right\|_2^2 + \gamma_{k-1}\hat{f}_{k-1}(z_k) \\
&= (1-\gamma_{k-1})\phi_{k-1}^* + \frac{\gamma_{k-1}^2(1-\gamma_{k-1})\eta_{k-1}}{2\eta_k^2}\|\nabla f(y_k)\|_2^2 + \gamma_{k-1}\hat{f}_{k-1}(z_k) \\
&= (1-\gamma_{k-1})\phi_{k-1}^* + \frac{\gamma_{k-1}^2}{2\eta_k}\|\nabla f(y_k)\|_2^2 + \gamma_{k-1}\hat{f}_{k-1}(z_k).
\end{aligned} \tag{4.15}$$

Since $\hat{f}_{k-1}(x) = f(y_k) + \nabla f(y_k)^T(x - y_k)$ and $z_k \leftarrow z_{k-1} - \frac{\gamma_{k-1}}{\eta_k}\nabla f(y_k)$,

$$\begin{aligned}
\hat{f}_{k-1}(z_k) &= f(y_k) + \nabla f(y_k)^T\left(z_{k-1} - \frac{\gamma_{k-1}}{\eta_k}\nabla f(y_k) - y_k\right) \\
&= \hat{f}_{k-1}(z_{k-1}) - \frac{\gamma_{k-1}^2}{\eta_k}\|\nabla f(y_k)\|_2^2.
\end{aligned} \tag{4.16}$$

We combine (4.15) and (4.16) to obtain (4.11). □

Why are estimating sequences useful? As it turns out, for any sequence $\{x_k\}$ obeying $f(x_k) \le \inf_x \phi_k(x)$, the convergence rate of $\{f(x_k)\}$ is given by the convergence rate of $\{\tilde{\eta}_k\}$.

COROLLARY 4.5. *If $\{\phi_k\}$ is an estimating sequence of f and if $f(x_k) \leq \inf_x \phi(x)$ for some sequence $\{x_k\}$, then*

$$f(x_k) - f^* \leq \tilde{\eta}_k \left(\phi_0(x^*) - f^*\right).$$

PROOF. Since $f(x_k) \leq \inf_x \phi_k(x)$,

$$\begin{aligned} f(x_k) - f^* \leq \phi_k(x^*) - f^* &\leq (1-\tilde{\eta}_k) f^* + \tilde{\eta}_k \phi_0(x^*) - f^* \\ &\leq \tilde{\eta}_k \left(\phi_0(x^*) - f^*\right). \end{aligned}$$

□

We show that Nesterov's choice of $\{\phi_k\}$ is an estimating sequence. First, we state two lemmas concerning the sequence $\{\gamma_k\}$. Their proofs are in the Appendix.

LEMMA 4.6. *The sequence $\{\gamma_k\}$ satisfies $\gamma_k^2 = \prod_{l=1}^k (1-\gamma_l)$.*

LEMMA 4.7. *At the k-th iteration, γ_k is at most $\frac{2}{k+2}$.*

LEMMA 4.8. *The functions $\{\phi_k\}$ form an estimating sequence of f :*

$$\phi_k(x) \leq (1-\tilde{\eta}_k) f(x) + \tilde{\eta}_k \phi_0(x),$$

where $\tilde{\eta}_k := \prod_{l=1}^k (1-\gamma_k)$. Further,

$$\tilde{\eta}_k \leq \frac{4}{(k+2)^2} \text{ for any } k \geq 0.$$

PROOF. Consider the sequence $\tilde{\eta}_0 = 0$, $\tilde{\eta}_k = \prod_{l=1}^k (1-\gamma_k)$. At x_0, clearly $\phi_0(x) \leq \phi_0(x)$. Since f is convex, $f(x) \geq \hat{f}_k(x)$, and

$$\phi_{k+1}(x) \leq (1-\gamma_k)\phi_k(x) + \gamma_k f(x).$$

By the induction hypothesis,

$$\begin{aligned} &\leq (1-\gamma_k)\left((1-\tilde{\eta}_{k-1}) f(x) + \tilde{\eta}_{k-1}\phi_0(x)\right) + \gamma_k f(x) \\ &= \left(1 - (1-\gamma_k)\tilde{\eta}_{k-1}\right) f(x) + (1-\gamma_k)\tilde{\eta}_{k-1}\phi_0(x) \\ &= (1-\tilde{\eta}_k) f(x) + \tilde{\eta}_k \phi_0(x). \end{aligned}$$

It remains to show $\tilde{\eta}_k \leq \frac{4}{(k+2)^2}$. By (4.12), $\tilde{\eta}_k = \prod_{l=1}^k (1-\gamma_l)$. Further, by Lemma (4.6),

$$\tilde{\eta}_k = \prod_{l=1}^k (1-\gamma_l) = \gamma_k^2. \tag{4.17}$$

We bound the right side by Lemma 4.7 to obtain the stated result. □

To invoke Corollary 4.5, we must first show $f(x_k) \leq \inf_x \phi_k(x)$. As we shall see, $\{\gamma_k\}$ and $\{y_k\}$ are carefully chosen to ensure $f(x_k) \leq \inf_x \phi_k(x)$. For now, assume $f(x_k) \leq \inf_x \phi_k(x)$. By (4.11) and the convexity of f,

$$\begin{aligned}
\phi_{k+1}^* &\geq (1-\gamma_k) f(x_k) + \gamma_k \hat{f}_k(z_k) - \frac{\gamma_k^2}{2\eta_{k+1}} \|\nabla f(y_{k+1})\|_2^2 \\
&\geq (1-\gamma_k) \hat{f}_k(x_k) + \gamma_k \hat{f}_k(z_k) - \frac{\gamma_k^2}{2\eta_{k+1}} \|\nabla f(y_{k+1})\|_2^2 \\
&= f(y_{k+1}) + \nabla f(y_{k+1})^T ((1-\gamma_k) x_k + \gamma_k z_k - y_{k+1}) \\
&\quad - \frac{\gamma_k^2}{2\eta_{k+1}} \|\nabla f(y_{k+1})\|_2^2.
\end{aligned}$$

By (4.9), the linear term is zero:

$$\phi_{k+1}^* \geq f(y_{k+1}) - \frac{\gamma_k^2}{2\eta_{k+1}} \|\nabla f(y_{k+1})\|_2^2.$$

To ensure $f(x_{k+1}) \leq \phi_{k+1}^*$, it suffices to ensure

$$f(x_{k+1}) \leq f(y_{k+1}) - \frac{\gamma_k^2}{2\eta_{k+1}} \|\nabla f(y_{k+1})\|_2^2,$$

which, as we shall see, is possible by taking a gradient step from y_{k+1}.

LEMMA 4.9. *Under the conditions of Theorem 4.1, the sequence* $\{x_k\}$ *given by* (4.10) *obeys* $f(x_k) \leq \phi_k^*$.

PROOF. We proceed by induction. At x_0, $f(x_0) = \phi_0^*$. At each iteration, x_{k+1} is a gradient step from y_{k+1}. By Lemma 4.2, the function value decreases by at least $\frac{1}{2L} \|\nabla f(y_{k+1})\|_2^2$:

$$f(x_{k+1}) \leq f(y_{k+1}) - \frac{1}{2L} \|\nabla f(y_{k+1})\|_2^2 = \hat{f}_k(y_{k+1}) - \frac{1}{2L} \|\nabla f(y_{k+1})\|_2^2,$$

where the second inequality follows from the definition of $\hat{f}_k$. By (4.9),

$$\begin{aligned}
\hat{f}_k(y_{k+1}) &= \hat{f}((1-\gamma_k) x_k + \gamma_k(z_k)) \\
&= (1-\gamma_k) \hat{f}_k(x_k) + \gamma_k \hat{f}_k(z_k).
\end{aligned}$$

Thus $f(x_{k+1})$ is further bounded by

$$\begin{aligned}
f(x_{k+1}) &\leq (1-\gamma_k) \hat{f}_k(x_k) + \gamma_k \hat{f}_k(z_k) - \frac{1}{2L} \|\nabla f(y_{k+1})\|_2^2 \\
&\leq (1-\gamma_k) f(x_k) + \gamma_k \hat{f}_k(z_k) - \frac{1}{2L} \|\nabla f(y_{k+1})\|_2^2 \\
&\leq (1-\gamma_k) \phi_k^* + \gamma_k \hat{f}_k(z_k) - \frac{1}{2L} \|\nabla f(y_{k+1})\|_2^2,
\end{aligned}$$

where the second inequality follows from the induction hypothesis. Substituting (4.14), we obtain

$$f(x_{k+1}) \le \phi_{k+1}^* + \left(\frac{\gamma_k^2}{2\eta_{k+1}} - \frac{1}{2L}\right)\|\nabla f(y_{k+1})\|_2^2.$$

By (4.12), $\eta_{k+1} = \frac{L}{2}\prod_{l=1}^{k}(1-\gamma_l)$. Thus

$$\frac{\gamma_k^2}{2\eta_{k+1}} - \frac{1}{2L} = \frac{1}{2L}\left(\frac{\gamma_k^2}{\prod_{l=1}^{k}(1-\gamma_l)} - 1\right).$$

The difference in parentheses is zero by Lemma 4.6. □

Thus the iterates $\{x_k\}$ obey $f(x_k) \le \inf_x \phi_k(x)$. We put the pieces together to obtain the convergence rate of Nesterov's 1983 method.

THEOREM 4.10. *Under the conditions of Theorem 4.1, the iterates* $\{x_k\}$ *generated by* (4.10) *satisfy*

$$f(x_k) - f^* \le \frac{4L}{(k+2)^2}\|x_0 - x^*\|_2^2.$$

PROOF. By Corollary 4.5 and Lemma 4.8, the function values obey

$$\begin{aligned} f(x_k) - f^* &\le \tilde{\eta}_k\left(\phi_0(x^*) - f^*\right) \\ &\le \frac{4}{(k+2)^2}\left(\phi_0(x^*) - f^*\right). \end{aligned}$$

Further, recalling $\phi_0(x) = f(x_0) + \frac{L}{2}\|x - x_0\|_2^2$, we obtain

$$\begin{aligned} f(x_k) - f^* &\le \frac{4}{(k+2)^2}\left(\frac{L}{2}\|x - x_0\|_2^2 + f(x_0) - f^*\right) \\ &\le \frac{4L}{(k+2)^2}\|x - x_0\|_2^2, \end{aligned}$$

where the second inequality follows from the strong smoothness of f. □

Finally, we check the iterates $\{(x_k, y_k, z_k)\}$ produced by forming auxiliary functions are the same as those generated by (4.10), (4.9), (4.8).

LEMMA 4.11. *The iterates* $\{(x_k, y_k, z_k)\}$ *generated by the estimating sequence are identical to those generated by* (4.10), (4.9), (4.8).

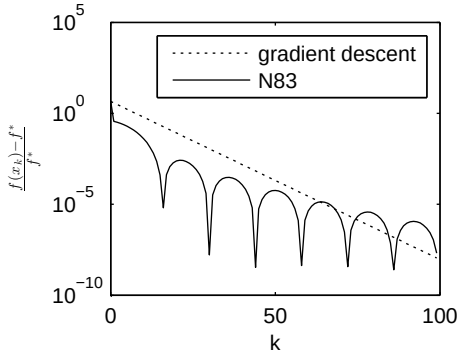

FIG 3. *Convergence of gradient descent and Nesterov's 1983 method on a strongly convex function. Like the heavy ball method, Nesterov's 1983 method generates function values that are not monotonically decreasing.*

PROOF. Since x_k, y_k are given by (4.10) and (4.9) when forming an estimating sequence, it suffices to check that (4.8) and (4.13) are identical.

$$\begin{aligned} z_{k+1} &\leftarrow x_k + \frac{1}{\gamma_k}(x_{k+1} - x_k) \\ &= x_k + \frac{1}{\gamma_k}\left(y_{k+1} - \frac{1}{L}\nabla f(y_{k+1}) - x_k\right) \\ &= x_k + \frac{1}{\gamma_k}\left((1-\gamma_k)x_k + \gamma_k z_k - \frac{1}{L}\nabla f(y_{k+1}) - x_k\right) \\ &= z_k - \frac{1}{\gamma_k L}\nabla f(y_{k+1}). \end{aligned}$$

It remains to show $\frac{1}{\gamma_k L} = \frac{\gamma_k}{\eta_{k+1}}$. By (4.12), $\eta_{k+1} = L\prod_{l=1}^{k}(1-\gamma_l)$. Thus

$$\frac{\gamma_k}{\eta_{k+1}} = \frac{\gamma_k}{L\prod_{l=1}^{k}(1-\gamma_l)} = \frac{1}{\gamma_k L},$$

where the second equality follows by Lemma 4.6. □

Nesterov's 1983 method, as stated, does not converge linearly when f is strongly convex. Figure 4.2 compares the convergence of gradient descent and Nesterov's 1983 method. It's possible to modify the method to attain linear convergence. The key idea is to form ϕ_k by convex combinations of ϕ_{k-1} and quadratic lower bounds of f given by (2.2). We refer to Nesterov (2004), Section 2.2 for details.

A more practical issue is choosing the step size in (4.10) when L is unknown. It is possible to incorporate a line search:

$$x_{k+1} \leftarrow y_{k+1} - \alpha_k \nabla f(y_{k+1}), \tag{4.18}$$

where α_k is chosen by a line search. To form a valid estimating sequence, the step sizes must decrease: $\alpha_k \leq \alpha_{k-1}$. If α_k is chosen by a backtracking line search, the initial trial step at the k-th iteration is set to α_{k-1}.

The key idea in Nesterov (1983) is forming an estimating sequence, and Nesterov's 1983 method is "merely" a special case. The development of accelerated first-order methods by forming estimating sequences remains an active area of research. Just recently,

1. Meng and Chen (2011) proposed a modification of Nesterov's 1983 method that further accelerates the speed of convergence.
2. Gonzaga and Karas (2013) proposed a variant of Nesterov's original method that adapts to unknown strong convexity.
3. O'Donoghue and Candés (2013) proposed a heuristic for resetting the momentum term to zero that improves the convergence rate.

To end on a practical note, the MATLAB package TFOCS by Becker, Candès and Grant (2011) implements some of the aforementioned methods.

4.3. *Optimality of accelerated gradient methods.* To wrap up the section, we pause and ask: are there methods that converge faster than accelerated gradient descent? In other words, given a convex function with Lipschitz gradient, what is the fastest convergence rate attainable by first-order methods? More precisely, a first-order method is one that chooses the k-th iterate in

$$x_0 + \operatorname{span}\{\nabla f(x_0), \ldots, \nabla f(x_{k-1})\} \text{ for } k = 1, 2, \ldots$$

As it turns out, the $O(\frac{1}{k^2})$ convergence rate attained by Nesterov's accelerated gradient method is optimal. That is, unless f has special extra structure (e.g. strong convexity), no method has better worst-case performance than Nesterov's 1983 method.

THEOREM 4.12. *There exists a convex and strongly smooth function* $f : \mathbf{R}^{2k+1} \to \mathbf{R}$ *such that*

$$f(x_k) - f^* \geq \frac{3L\|x_0 - x^*\|_2^2}{32(k+1)^2},$$

where $\{x_l\}$ *obey* $x_l \in x_0 + \operatorname{span}(\{\nabla f(x_0), \ldots, \nabla f(x_{l-1})\})$ *for any* $l \in [\,k\,]$.

PROOF SKETCH. It suffices to exhibit a function that is hard to optimize. More precisely, we construct a function and initial guess such that any first-order method has at best $O(\frac{1}{k^2})$ convergence rate when applied to the problem. Consider the convex quadratic function

$$f(x) = \frac{L}{4}\left(\frac{1}{2}x_1^2 + \frac{1}{2}\sum_{i=1}^{2k}(x_i - x_{i+1})^2 + \frac{1}{2}x_{2k+1}^2 - x_1\right).$$

In matrix form, $f(x) = \frac{L}{4}\left(\frac{1}{2}x^T A x - e_1^T x\right)$, where

$$A = \begin{bmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 2 \end{bmatrix} \in \mathbf{R}^{(2k+1)\times(2k+1)}.$$

It's possible to show

1. f is strongly smooth with constant L,
2. its minimum $\frac{L}{8}\left(\frac{1}{2k+2} - 1\right)$ is attained at $\left[x^*\right]_i = 1 - \frac{i}{2k+2}$.
3. $\mathcal{K}_l[f] := \operatorname{span}(\{\nabla f(x_0), \ldots, \nabla f(x_{l-1})\})$ is $\operatorname{span}(\{e_1, \ldots, e_l\})$.

If we initialize $x_0 \leftarrow 0$, then at the k-th iteration,

$$f(x_k) \geq \inf_{x \in \mathcal{K}_k[f]} f(x) = \frac{L}{8}\left(\frac{1}{k+1} - 1\right).$$

Thus

$$\frac{f(x_k) - f^*}{\|x_0 - x^*\|_2^2} \geq \frac{\frac{L}{8}\left(\frac{1}{k+1} - \frac{1}{2k+2}\right)}{\frac{1}{3}(2k+2)} = \frac{3L}{32(k+1)^2}.$$

□

5. Projected gradient descent. To wrap up, we consider a simple modification of gradient descent for constrained optimization: *projected gradient descent.* At each iteration, we take a gradient step and project back onto the feasible set:

$$x_{k+1} \leftarrow P_C(x_k - \alpha_k \nabla f(x_k)), \tag{5.1}$$

where $P_C(x) := \arg\min_{y \in C} \frac{1}{2}\|x - y\|_2^2$ is the projection of x onto C. For many sets (e.g. subspace, (symmetric) cones, norm balls etc.), it is possible

to project onto the set efficiently, making projected gradient descent a viable choice for optimization over the set.

Despite the requirement of being able to project efficiently onto the feasible set, projected gradient descent is quite general. Consider the (generic) nonlinear optimization problem

$$\begin{array}{ll} \underset{x\in\mathbf{R}^n}{\text{minimize}} & f(x) \\ \text{subject to} & l \le c(x) \le u. \end{array} \tag{5.2}$$

By introducing slack variables, (5.2) is equivalent to

$$\begin{array}{ll} \underset{s\in\mathbf{R}^m, x\in\mathbf{R}^n}{\text{minimize}} & f(x) \\ \text{subject to} & c(x) = s, \quad l \le s \le u. \end{array}$$

The *bound-constrained Lagrangian approach* (BCL) solves a sequence of bound-constrained augmented Lagrangian subproblems of the form

$$\begin{array}{ll} \underset{s\in\mathbf{R}^m, x\in\mathbf{R}^n}{\text{minimize}} & f(x) + \lambda_k^T c(x) + \dfrac{\rho_k}{2}\left\|c(x) - s\right\|_2^2 \\ \text{subject to} & l \le s \le u. \end{array}$$

Since projection onto rectangles is efficient, projected gradient descent may be used to solve the BCL subproblems.

Given the similarity of projected gradient descent to gradient descent, it is no suprise that the convergence rate of projected gradient descent is the same as that of its counterpart. For now, fix a step size α and consider projected gradient descent as a fixed point iteration:

$$x_{k+1} \leftarrow P_C(G_\alpha(x_k)),$$

where G_α is the gradient step. Since the optimum x^* is a fixed point of $P_C \circ G_\alpha$, the iterates converge if $P_C \circ G_\alpha$ is a contraction.

LEMMA 5.1. *The projection onto a convex set $C \subset \mathbf{R}^n$ is non-expansive:*

$$\|P_C(y) - P_C(x)\|_2^2 \le (x - y)^T (P_C(y) - P_C(x)).$$

PROOF. The optimality condition of the constrained optimization problem defining the projection onto C is

$$(x - P_C(x))^T (y - P_C(x)) \le 0 \text{ for any } y \in C.$$

For any two points x, y and their projections onto C, we have

$$(x - P_C(x))^T (P_C(y) - P_C(x)) \leq 0$$
$$(y - P_C(y))^T (P_C(x) - P_C(y)) \leq 0.$$

We add the two inequalities and rearrange to obtain

$$\|P_C(y) - P_C(x)\|_2^2 \leq (x - y)^T (P_C(y) - P_C(x)).$$

We apply the Cauchy-Schwartz inequality to obtain the stated result. □

Since P_C is non-expansive,

$$\|P_C(G_\alpha(x)) - P_C(G_\alpha(y))\|_2 \leq \|G_\alpha(x) - G_\alpha(y)\|_2.$$

Thus, as long as α is chosen to ensure G_α is a contraction, $P_C \circ G_\alpha$ is also a contraction. The rest of the story is the same as that of gradient descent: when f is strongly convex and strongly smooth, projected gradient descent converges linearly. That is, it requires at most $O(\kappa \log(\frac{1}{\epsilon}))$ iterations to obtain an ϵ-accurate solution.

When f is not strongly convex (but convex and strongly smooth), the story is also similar to that of gradient descent. However, the proof is more involved. First, we show projected gradient descent is a descent method.

LEMMA 5.2. *When*

1. $f : \mathbf{R}^n \to \mathbf{R}$ *is convex and strongly smooth with constant* L
2. $C \subset \mathbf{R}^n$ *is convex,*

projected gradient descent with step size $\frac{1}{L}$ *satisfies*

$$f(x_{k+1}) \leq f(x_k) - \frac{L}{2}\|x_{k+1} - x_k\|_2^2.$$

PROOF. Since f is strongly smooth, we have

$$f(x_{k+1}) \leq f(x_k) + \nabla f(x_k)^T (x_{k+1} - x_k) + \frac{L}{2}\|x_{k+1} - x_k\|_2^2. \tag{5.3}$$

The point x_{k+1} is the projection of $x_k - \frac{1}{L}\nabla f(x_k)$ onto C. Thus

$$(x_k - \frac{1}{L}\nabla f(x_k) - x_{k+1})^T (y - x_{k+1}) \leq 0 \text{ for any } y \in C. \tag{5.4}$$

By setting $y = x_k$ in (5.4) and rearranging, we obtain

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

We substitute the bound into (5.3) to obtain the stated result. □

We put the pieces together to show projected gradient descent requires at most $O(\frac{L}{\epsilon})$ iterations to obtain an ϵ-suboptimal point.

THEOREM 5.3. *Under the conditions of Lemma 5.2, projected gradient descent with step size* $\frac{1}{L}$ *satisfies*

$$f(x_k) - f^* \leq \frac{L}{2k}\|x^* - x_0\|_2^2.$$

PROOF. By setting $y = x^*$ in (5.4) (and rearranging), we obtain

$$\nabla f(x_k)^T(x_{k+1} - x^*) \leq L(x_{k+1} - x_k)^T(x^* - x_{k+1}).$$

Substituting the bound into (5.3), we have

$$f(x_{k+1}) \leq f(x_k) + L(x_{k+1} - x_k)^T(x^* - x_{k+1}) + \frac{L}{2}\|x_{k+1} - x_k\|_2^2.$$

We complete the square to obtain

$$\begin{aligned} f(x_{k+1}) &\leq f(x_k) + L\|x^* - x_{k+1} + x_{k+1} - x_k\|_2^2 - \frac{L}{2}\|x^* - x_{k+1}\|_2^2 \\ &= f(x_k) + \frac{L}{2}\left(\|x^* - x_k\|_2^2 - \|x^* - x_{k+1}\|_2^2\right). \end{aligned}$$

Summing over iterations, we obtain

$$\begin{aligned} \sum_{l=1}^{k} f(x_l) - f^* &\leq \frac{L}{2}\sum_{l=1}^{k} \|x^* - x_{k-1}\|_2^2 - \|x^* - x_k\|_2^2 \\ &= \frac{L}{2}\left(\|x^* - x_0\|_2^2 - \|x^* - x_k\|_2^2\right) \\ &\leq \frac{L}{2}\|x^* - x_0\|_2^2. \end{aligned}$$

By Lemma 5.2, $\{f(x_k)\}$ is decreasing. Thus

$$f(x_k) - f^* \leq \frac{1}{k}\sum_{l=1}^{k} f(x_l) - f^* \leq \frac{L}{2k}\|x^* - x_0\|_2^2.$$

□

When the strong smoothness constant is unknown, it is possible to choose step sizes by a projected backtracking line search. We

1. start with some initial trial step size $\alpha_k \leftarrow \alpha_{\text{init}}$

2. decrease the trial step size geometrically: $\alpha_k \leftarrow \frac{\alpha_k}{2}$
3. continue until α_k satisfies a sufficient descent condition:

$$f(x_{k+1}) \leq f(x_k) + \nabla f(x_k)^T(x_{k+1} - x_k) + \frac{L}{2}\|x_{k+1} - x_k\|_2^2, \tag{5.5}$$

where $x_{k+1} \leftarrow P_C(x_k - \alpha_k \nabla f(x_k))$.

Thus the "line search" backtracks along the projection of the search ray onto the feasible set. It is possible to show projected gradient descent with a projected backtracking line search preserves the $O(\frac{1}{k})$ convergence rate of projected gradient descent with step size $\frac{1}{L}$.

APPENDIX

PROOF OF LEMMA 4.6. We proceed by induction. Recall $\{\gamma_k\}$ is generated by the recursion (4.7). Since $\gamma_0 \leftarrow 1$, we automatically have

$$\gamma_1^2 = (1 - \gamma_1).$$

By the induction hypothesis $\gamma_{k-1}^2 = \prod_{l=1}^{k-1}(1 - \gamma_l)$ and (4.7),

$$\gamma_k^2 = (1 - \gamma_k)\gamma_{k-1}^2 = \prod_{l=1}^{k}(1 - \gamma_l).$$

□

PROOF OF LEMMA 4.7. We consider the gap between successive $\frac{1}{\gamma_k}$:

$$\frac{1}{\gamma_{l+1}} - \frac{1}{\gamma_l} = \frac{\gamma_l - \gamma_{l+1}}{\gamma_l \gamma_{l+1}} = \frac{\gamma_l^2 - \gamma_{l+1}^2}{\gamma_l \gamma_{l+1}(\gamma_l + \gamma_{l+1})}.$$

By (4.7), we have

$$\frac{1}{\gamma_{l+1}} - \frac{1}{\gamma_l} = \frac{\gamma_l^2 - \gamma_l^2(1 - \gamma_{l+1})}{\gamma_l \gamma_{l+1}(\gamma_l + \gamma_{l+1})} = \frac{\gamma_l^2 \gamma_{l+1}}{\gamma_l \gamma_{l+1}(\gamma_l + \gamma_{l+1})}.$$

It is easy to show the sequence $\{\gamma_l\}$ is decreasing. Thus

$$\frac{1}{\gamma_{l+1}} - \frac{1}{\gamma_l} \geq \frac{\gamma_l^2 \gamma_{l+1}}{\gamma_l \gamma_{l+1}(2\gamma_l)} = \frac{1}{2}.$$

Summing the inequalities for $l = 0, 1, \ldots, k-1$, we obtain

$$\frac{1}{\gamma_k} - \frac{1}{\gamma_0} \geq \frac{k}{2}.$$

We recall $\gamma_0 \leftarrow 1$ to deduce $\frac{1}{\gamma_k} \geq 1 + \frac{k}{2}$, which implies the second bound. □

REFERENCES

Beck, A. and Teboulle, M. (2009). Gradient-based algorithms with applications to signal recovery. *Convex Optimization in Signal Processing and Communications*.

Becker, S. R., Candès, E. J. and Grant, M. C. (2011). Templates for convex cone problems with applications to sparse signal recovery. *Mathematical Programming Computation* **3** 165–218.

Gonzaga, C. C. and Karas, E. W. (2013). Fine tuning Nesterov's steepest descent algorithm for differentiable convex programming. *Mathematical Programming* **138** 141–166.

Meng, X. and Chen, H. (2011). Accelerating Nesterov's method for strongly convex functions with Lipschitz gradient. *arXiv preprint arXiv:1109.6058*.

Nesterov, Y. (1983). A method of solving a convex programming problem with convergence rate $O(\frac{1}{k^2})$. *Soviet Mathematics Doklady* **27** 372–376.

Nesterov, Y. (2004). *Introductory Lectures on Convex Optimization* **87**. Springer Science & Business Media.

O'Donoghue, B. and Candés, E. (2013). Adaptive restart for accelerated gradient schemes. *Foundations of Computational Mathematics* 1–18.

Polyak, B. T. (1964). Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics* **4** 1–17.

Yuekai Sun
Stanford, California
April 29, 2015
E-mail: yuekai@stanford.edu