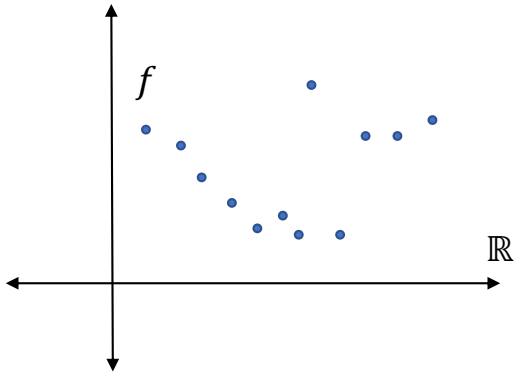
Introduction to Optimization Theory

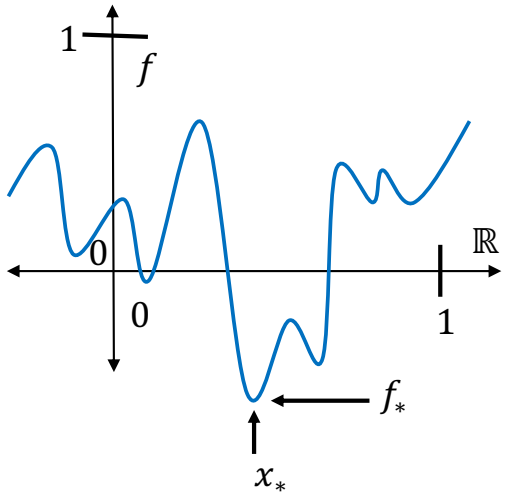
Lecture #7 - 10/6/20

MS&E 213 / CS 2690



Aaron Sidford

sidford@stanford.edu



Plan for Today

Recap

- Accelerated Gradient Descent (AGD)

Proof

- Approximately optimal AGD for smooth strongly convex functions.

Extensions

- Non-strongly convex
- Optimal complexity
- Momentum

Thursday

Generalizations and applications

Recap

Problem

$$\min_{x \in \mathbb{R}^n} f(x)$$

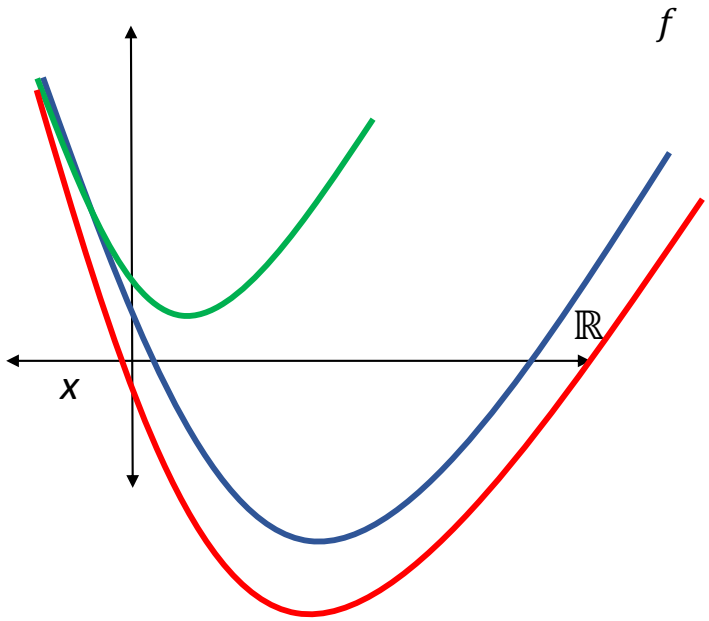
Regularity	Oracle	Goal	Algorithm	Iterations
$n = 1, f(x) \in [0,1], x_* \in [0,1]$	value	½-optimal	anything	∞
$n = 1, x_* \in [0,1]$, L-Lipschitz	value	ϵ-optimal	ϵ-net	$\Theta(L/\epsilon)$
$x_* \in [0,1]$, L-Lipschitz in $\|\cdot\|_\infty$	value	ϵ-optimal	ϵ-net	$\left(\Theta(L/\epsilon)\right)^n$
L-smooth and bounded	value, gradient	ϵ-optimal	ϵ-net	exponential
L-smooth	gradient	ϵ-critical	gradient descent	$O\left(\frac{L(f(x_0) - f_*)}{\epsilon^2}\right)$
L-smooth μ-strongly convex	gradient	ϵ-optimal	gradient descent	$O\left(\frac{L}{\mu}\log\left(\frac{f(x_0) - f_*}{\epsilon}\right)\right)$
L-smooth convex	gradient	ϵ-optimal	gradient descent	$O\left(\frac{L\|x_0 - x_*\|_2^2}{\epsilon}\right)$

Today: prove and discuss improvements to $O\left(\sqrt{\frac{L}{\mu}}\log\left(\frac{f(x_0)-f_*}{\epsilon}\right)\right)$ ***and*** $O\left(\sqrt{\frac{L\|x_0-x_*\|_2^2}{\epsilon}}\right)$

Recap

Theorem: $f: \mathbb{R}^n \to \mathbb{R}$ is L-smooth and μ-strongly convex (with respect to $\|\cdot\|_2$) if and only if the following hold for all x, y

- $f(y) \leq \boldsymbol{U_x(y)} \stackrel{\text{def}}{=} f(x) + \nabla f(x)^\top (y - x) + \frac{L}{2}\|y - x\|_2^2$
- $f(y) \geq \boldsymbol{L_x(y)} \stackrel{\text{def}}{=} f(x) + \nabla f(x)^\top (y - x) + \frac{\mu}{2}\|y - x\|_2^2$



Goal #1: improve

$$O\left(\frac{L}{\mu}\log\left(\frac{f(x_0)-f_*}{\epsilon}\right)\right) \text{ to } \sim O\left(\sqrt{\frac{L}{\mu}}\log\left(\frac{f(x_0)-f_*}{\epsilon}\right)\right)$$

Approach

- *Are many to choose from*
- *This one is intuitive and mechanical and will let us touch on ideas of other proof*
- *Will lose a logarithmic factor and will explain how to remove*

Approach

- Maintain point ($x_k \in \mathbb{R}^n$)
- Maintain lower bound
  - $L_k : \mathbb{R}^n \to \mathbb{R}$ s.t.
  - $L_k(x) \leq f(x)$ for all $x \in \mathbb{R}^n$
- Update both in each iteration

Progress Measure

- $f(x_k) - \min_{x \in \mathbb{R}^n} L_k(x)$
- $\geq f(x_k) - L_k(x_*) = f(x_k) - f_*$

Tools: Quadratic Lower Bounds

Lemma 1: $L_y(x) = f(y) + \nabla f(y)^\top (x-y) + \frac{L}{2}\|x-y\|_2^2 = \psi_y + \frac{\mu}{2}\|x-v_y\|_2^2$
for $\psi_y = f(y) - \frac{1}{2\mu}\|\nabla f(y)\|_2^2$ and $v_y = y - \frac{1}{\mu}\nabla f(y)$.

Lemma 2: If $f_1, f_2 \colon \mathbb{R}^n \to \mathbb{R}$ are defined for all $x \in \mathbb{R}^n$ by

$$f_1(x) = \psi_1 + \frac{\mu}{2}\|x-v_1\|_2^2 \quad \text{and} \quad f_2(x) = \psi_2 + \frac{\mu}{2}\|x-v_2\|_2^2$$

Then for all $\alpha \in [0,1]$ we have

$$f_\alpha(x) = \alpha \cdot f_1(x) + (1-\alpha)\cdot f_2(x) = \psi_\alpha + \frac{\mu}{2}\|x-v_\alpha\|_2^2$$

Where

- $v_\alpha = \alpha \cdot v_1 + (1-\alpha)\cdot v_2$
- $\psi_\alpha = \alpha \cdot \psi_1 + (1-\alpha)\cdot \psi_2 + \frac{\mu}{2}\alpha(1-\alpha)\|v_1 - v_2\|_2^2$

Accelerated Gradient Descent (AGD)

- Initial $x_0 \in \mathbb{R}^n$, $L_0(x) = \psi_0 + \frac{\mu}{2}\|x - v_0\|_2^2$ s.t. $f(x) \geq L_0(x)$ for all x
- Repeat for $k = 0,1,2, \ldots$
  - $y_k = \alpha \cdot x_k + (1-\alpha) \cdot v_k$ where $\alpha \in [0,1]$
  - $L_{y_k}(x) = f(y_k) + \nabla f(y_k)^\top (x - y_k) + \frac{\mu}{2}\|x - y_k\|_2^2$
  - $L_{k+1}(x) = \psi_{k+1} + \frac{\mu}{2}\|x - v_{k+1}\|_2^2 = \beta L_k(x) + (1-\beta) L_{y_{k+1}}(x)$ where $\beta \in [0,1]$
  - $x_{k+1} = y_k - \frac{1}{L}\nabla f(y_k)$

Theorem: $L_k(x) \geq f(x)$ for all $k \geq 0$ and $x \in \mathbb{R}^n$. If $\kappa = \frac{L}{\mu}$, $\alpha = \frac{\sqrt{\kappa}}{\sqrt{\kappa}+1}$, and $\beta = 1 - \kappa^{-1/2}$, then

$$f(x_{k+1}) - \psi_{k+1} \leq \left(1 - \frac{1}{\sqrt{\kappa}}\right)[f(x_k) - \psi_k]$$

and $\sim\sqrt{\kappa}$ iterations suffices

Proof?

Plan for Today

Recap

- Accelerated Gradient Descent (AGD)

Proof

- Approximately optimal AGD for smooth strongly convex functions.

Extensions

- Non-strongly convex
- Optimal complexity
- Momentum

Thursday

Generalizations and applications

Accelerated Gradient Descent (AGD)

- Initial $x_0 \in \mathbb{R}^n$, $L_0(x) = \psi_0 + \frac{\mu}{2}\|x - v_0\|_2^2$ s.t. $f(x) \geq L_0(x)$ for all x
- Repeat for $k = 0,1,2, \ldots$
  - $y_k = \alpha \cdot x_k + (1-\alpha) \cdot v_k$ where $\alpha \in [0,1]$
  - $L_{y_k}(x) = f(y_k) + \nabla f(y_k)^\top (x - y_k) + \frac{\mu}{2}\|x - y_k\|_2^2$
  - $L_{k+1}(x) = \psi_{k+1} + \frac{\mu}{2}\|x - v_{k+1}\|_2^2 = \beta L_k(x) + (1-\beta) L_{y_{k+1}}(x)$ where $\beta \in [0,1]$
  - $x_{k+1} = y_k - \frac{1}{L}\nabla f(y_k)$

$v_{k+1} = \beta v_k + (1-\beta)\left[y_k - \frac{1}{\mu}\nabla f(y_k\right.$

Theorem: $L_k(x) \geq f(x)$ for all $k \geq 0$ and $x \in \mathbb{R}^n$. If $\kappa = \frac{L}{\mu}$, $\alpha = \frac{\sqrt{\kappa}}{\sqrt{\kappa}+1}$, and $\beta = 1 - \kappa^{-1/2}$, then

$$f(x_{k+1}) - \psi_{k+1} \leq \left(1 - \frac{1}{\sqrt{\kappa}}\right)[f(x_k) - \psi_k]$$

and $\sim\sqrt{\kappa}$ iterations suffices

Analysis?

Some Intuition

- Initial $x_0 \in \mathbb{R}^n$, $L_0(x) = \psi_0 + \frac{\mu}{2}\|x - v_0\|_2^2$ s.t. $f(x) \geq L_0(x)$ for all x
- Repeat for $k = 0,1,2, \ldots$
  - $y_k = \alpha \cdot x_k + (1-\alpha) \cdot v_k$ where $\alpha = \frac{\sqrt{\kappa}}{\sqrt{\kappa}+1}$ and $\kappa = \frac{L}{\mu}$
  - $v_{k+1} = \beta v_k + (1-\beta)\left[y_k - \frac{1}{\mu}\nabla f(y_k)\right]$ *where* $\beta = 1 - \frac{1}{\sqrt{\kappa}}$
  - $x_{k+1} = y_k - \frac{1}{L}\nabla f(y_k)$

Note

- $\uparrow \kappa \Rightarrow \uparrow \alpha$ (i.e. the more use gradient point)
- $\uparrow \kappa \Rightarrow \uparrow \beta$ (i.e. the less use lower bound)
- $\uparrow \kappa \Rightarrow \uparrow (1-\beta)/\mu$ (i.e. the bigger the "gradient step" for v_{k+1})

Analysis?

Proof Plan

- $y_k = \alpha \cdot x_k + (1-\alpha) \cdot v_k$
- $L_{y_k}(x) = f(y_k) = \nabla f(y_k)^\top (x - y_k) + \frac{\mu}{2}\|x - y_k\|_2^2$
- $L_{k+1}(x) = \psi_{k+1} + \frac{\mu}{2}\|x - v_{k+1}\|_2^2 = \beta L_k(x) + (1-\beta) L_{y_{k+1}}(x)$
- $x_{k+1} = y_k - \frac{1}{L}\nabla f(y_k)$

Theorem: $L_k(x) \geq f(x)$ for all $k \geq 0$ and $x \in \mathbb{R}^n$ and if $\kappa = \frac{L}{\mu}$, $\alpha = \frac{\sqrt{\kappa}}{\sqrt{\kappa}+1}$, and $\beta = 1 - \frac{1}{\sqrt{\kappa}}$, then

$$f(x_{k+1}) - \psi_{k+1} \leq \left(1 - \frac{1}{\sqrt{\kappa}}\right)[f(x_k) - \psi_k]$$

and $\sim\sqrt{\kappa}$ iterations suffices

Plan (*since $L_k(x) \geq f(x)$ fact is immediate*)

- Upper bound $f(x_{k+1})$ *(gradient descent step)*
- Lower bound ψ_k *(lower bound combination analysis)*
- Leverage choice of y_k *(algebra)*
- Pick α and β so everything cancels *(more algebra)*

Upper bound

- $y_k = \alpha \cdot x_k + (1-\alpha) \cdot v_k$
- $L_{y_k}(x) = f(y_k) = \nabla f(y_k)^\top (x - y_k) + \frac{\mu}{2} \|x - y_k\|_2^2$
- $L_{k+1}(x) = \psi_{k+1} + \frac{\mu}{2} \|x - v_{k+1}\|_2^2 = \beta L_k(x) + (1-\beta) L_{y_{k+1}}(x)$
- $x_{k+1} = y_k - \frac{1}{L} \nabla f(y_k)$

- $f(x_{k+1}) \leq ???$
- Gradient descent!
- $f(x_{k+1}) \leq f(y_k) - \frac{1}{2L} \|\nabla f(y_k)\|_2^2$

Proof Plan

- $y_k = \alpha \cdot x_k + (1-\alpha) \cdot v_k$
- $L_{y_k}(x) = f(y_k) = \nabla f(y_k)^\top (x - y_k) + \frac{\mu}{2}\|x - y_k\|_2^2$
- $L_{k+1}(x) = \psi_{k+1} + \frac{\mu}{2}\|x - v_{k+1}\|_2^2 = \beta L_k(x) + (1-\beta) L_{y_{k+1}}(x)$
- $x_{k+1} = y_k - \frac{1}{L}\nabla f(y_k)$

Theorem: $L_k(x) \geq f(x)$ for all $k \geq 0$ and $x \in \mathbb{R}^n$ and if $\kappa = \frac{L}{\mu}$, $\alpha = \frac{\sqrt{\kappa}}{\sqrt{\kappa}+1}$, and $\beta = 1 - \frac{1}{\sqrt{\kappa}}$, then

$$f(x_{k+1}) - \psi_{k+1} \leq \left(1 - \frac{1}{\sqrt{\kappa}}\right)[f(x_k) - \psi_k]$$

and $\sim\sqrt{\kappa}$ iterations suffices

Plan (*since $L_k(x) \geq f(x)$ fact is immediate*)

- Upper bound: $f(x_{k+1}) \leq f(y_k) - \frac{1}{2L}\|\nabla f(y_k)\|_2^2$
- Lower bound ψ_k *(lower bound combination analysis)*
- Leverage choice of y_k *(algebra)*
- Pick α and β so everything cancels *(more algebra)*

Lower Bound

- $y_k = \alpha \cdot x_k + (1-\alpha) \cdot v_k$
- $L_{y_k}(x) = f(y_k) = \nabla f(y_k)^\top (x - y_k) + \frac{\mu}{2}\|x - y_k\|_2^2$
- $L_{k+1}(x) = \psi_{k+1} + \frac{\mu}{2}\|x - v_{k+1}\|_2^2 = \beta L_k(x) + (1-\beta) L_{y_{k+1}}(x)$
- $x_{k+1} = y_k - \frac{1}{L}\nabla f(y_k)$

- Apply Tool #1
  - $L_{y_k}(x) = \psi_{y_k} + \frac{\mu}{2}\|x - v_{y_k}\|_2^2$
  - $\psi_{y_k} = f(y_k) - \frac{1}{2\mu}\|\nabla f(y_k)\|_2^2$ and $v_{y_k} = y_k - \frac{1}{\mu}\nabla f(y_k)$.
- Apply Tool #2
  - $\psi_{k+1} = \beta\psi_k + (1-\beta)\psi_{y_k} + \frac{\mu}{2}\beta(1-\beta)\left\|v_k - v_{y_k}\right\|_2^2$
- Algebra
  - $\left\|v_k - v_{y_k}\right\|_2^2 = \|v_k - y_k\|_2^2 + \frac{2}{\mu}\nabla f(y_k)^\top (v_k - y_k) + \frac{1}{\mu^2}\|\nabla f(y_k)\|_2^2$
- More algebra
  - $\psi_{k+1} \geq \beta\psi_k + (1-\beta)\left[f(y_k) - \frac{1-\beta}{2\mu}\|\nabla f(y_k)\|_2^2 + \beta\nabla f(y_k)^\top (v_k - y_k)\right]$

Proof Plan

- $y_k = \alpha \cdot x_k + (1-\alpha) \cdot v_k$
- $L_{y_k}(x) = f(y_k) = \nabla f(y_k)^\top (x - y_k) + \frac{\mu}{2} \|x - y_k\|_2^2$
- $L_{k+1}(x) = \psi_{k+1} + \frac{\mu}{2} \|x - v_{k+1}\|_2^2 = \beta L_k(x) + (1-\beta) L_{y_{k+1}}(x)$
- $x_{k+1} = y_k - \frac{1}{L} \nabla f(y_k)$

Theorem: $L_k(x) \geq f(x)$ for all $k \geq 0$ and $x \in \mathbb{R}^n$ and if $\kappa = \frac{L}{\mu}$, $\alpha = \frac{\sqrt{\kappa}}{\sqrt{\kappa}+1}$, and $\beta = 1 - \frac{1}{\sqrt{\kappa}}$, then

$$f(x_{k+1}) - \psi_{k+1} \leq \left(1 - \frac{1}{\sqrt{\kappa}}\right) [f(x_k) - \psi_k]$$

and $\sim\sqrt{\kappa}$ iterations suffices

Plan (*since $L_k(x) \geq f(x)$ fact is immediate*)

- Upper bound: $f(x_{k+1}) \leq f(y_k) - \frac{1}{2L} \|\nabla f(y_k)\|_2^2$
- Lower bound: $\psi_{k+1} \geq \beta \psi_k + (1-\beta) \left[f(y_k) - \frac{1-\beta}{2\mu} \|\nabla f(y_k)\|_2^2 + \beta \nabla f(y_k)^\top (v_k - y_k) \right]$
- Leverage choice of y_k *(algebra)*
- Pick α and β so everything cancels *(more algebra)*

Choice of y_k

- $y_k = \alpha \cdot x_k + (1-\alpha) \cdot v_k$
- $L_{y_k}(x) = f(y_k) = \nabla f(y_k)^\top (x - y_k) + \frac{\mu}{2}\|x - y_k\|_2^2$
- $L_{k+1}(x) = \psi_{k+1} + \frac{\mu}{2}\|x - v_{k+1}\|_2^2 = \beta L_k(x) + (1-\beta) L_{y_{k+1}}(x)$
- $x_{k+1} = y_k - \frac{1}{L}\nabla f(y_k)$

- **Goal**
  - Lower bound $\nabla f(y_k)^\top (v_k - y_k)$
- **Note**
  - $(1-\alpha)(v_k - y_k) + \alpha(x_k - y_k) = 0$
  - $v_k - y_k = \left(\frac{\alpha}{1-\alpha}\right)(y_k - x_k)$
  - *(note there is an $\alpha \in [0,1]$ s.t. $\frac{\alpha}{1-\alpha} = \gamma$ for all $\gamma > 0$)*
- **Convexity**
  - $f(x_k) \geq f(y_k) + \nabla f(y_k)^\top (x_k - y_k)$
  - *(note, this is the first time we have used convexity between two points where one of the points is not x_*)*
- **Algebra**
  - $\nabla f(y_k)^\top (v_k - y_k) \geq \left(\frac{\alpha}{1-\alpha}\right)[f(y_k) - f(x_k)]$

Proof Plan

- $y_k = \alpha \cdot x_k + (1-\alpha) \cdot v_k$
- $L_{y_k}(x) = f(y_k) = \nabla f(y_k)^\top (x - y_k) + \frac{\mu}{2}\|x - y_k\|_2^2$
- $L_{k+1}(x) = \psi_{k+1} + \frac{\mu}{2}\|x - v_{k+1}\|_2^2 = \beta L_k(x) + (1-\beta) L_{y_{k+1}}(x)$
- $x_{k+1} = y_k - \frac{1}{L}\nabla f(y_k)$

Theorem: $L_k(x) \geq f(x)$ for all $k \geq 0$ and $x \in \mathbb{R}^n$ and if $\kappa = \frac{L}{\mu}$, $\alpha = \frac{\sqrt{\kappa}}{\sqrt{\kappa}+1}$, and $\beta = 1 - \frac{1}{\sqrt{\kappa}}$, then

$$f(x_{k+1}) - \psi_{k+1} \leq \left(1 - \frac{1}{\sqrt{\kappa}}\right)[f(x_k) - \psi_k]$$

and $\sim\sqrt{\kappa}$ iterations suffices

Plan (*since $L_k(x) \geq f(x)$ fact is immediate*)

- Upper bound: $f(x_{k+1}) \leq f(y_k) - \frac{1}{2L}\|\nabla f(y_k)\|_2^2$
- Lower bound: $\psi_{k+1} \geq \beta\psi_k + (1-\beta)\left[f(y_k) - \frac{1-\beta}{2\mu}\|\nabla f(y_k)\|_2^2 + \beta\nabla f(y_k)^\top (v_k - y_k)\right]$
- Choice of y_k: $\nabla f(y_k)^\top (v_k - y_k) \geq \left(\frac{\alpha}{1-\alpha}\right)[f(y_k) - f(x_k)]$
- Pick α and β so everything cancels *(more algebra)*

Algebra

- $y_k = \alpha \cdot x_k + (1-\alpha) \cdot v_k$
- $L_{y_k}(x) = f(y_k) = \nabla f(y_k)^\top (x - y_k) + \frac{\mu}{2} \|x - y_k\|_2^2$
- $L_{k+1}(x) = \psi_{k+1} + \frac{\mu}{2} \|x - v_{k+1}\|_2^2 = \beta L_k(x) + (1-\beta) L_{y_{k+1}}(x)$
- $x_{k+1} = y_k - \frac{1}{L} \nabla f(y_k)$

So Far (*since $L_k(x) \geq f(x)$ fact is immediate*)

- Upper bound: $f(x_{k+1}) \leq f(y_k) - \frac{1}{2L} \|\nabla f(y_k)\|_2^2$
- Lower bound: $\psi_{k+1} \geq \beta \psi_k + (1-\beta) \left[f(y_k) - \frac{1-\beta}{2\mu} \|\nabla f(y_k)\|_2^2 + \beta \nabla f(y_k)^\top (v_k - y_k) \right]$
- Choice of y_k: $\nabla f(y_k)^\top (v_k - y_k) \geq \left(\frac{\alpha}{1-\alpha} \right) [f(y_k) - f(x_k)]$

Rearranging

- $$\begin{aligned} f(x_{k+1}) - \psi_{k+1} \leq\ & f(y_k) - \frac{1}{2L} \|\nabla f(y_k)\|_2^2 \\ & -\beta \psi_k - (1-\beta) \left[f(y_k) - \frac{1-\beta}{2\mu} \|\nabla f(y_k)\|_2^2 \right] \\ & -\beta (1-\beta) \left(\frac{\alpha}{1-\alpha} \right) [f(y_k) - f(x_k)] \\ =\ & \beta \left[\left(\frac{\alpha(1-\beta)}{1-\alpha} \right) f(x_k) - \psi_k \right] + \beta \left[1 - \left(\frac{\alpha(1-\beta)}{1-\alpha} \right) \right] f(y_k) + \frac{1}{2} \left[\frac{(1-\beta)^2}{\mu} - \frac{1}{L} \right] \|\nabla f(y_k)\|_2^2 \end{aligned}$$

Cancellations

- $y_k = \alpha \cdot x_k + (1-\alpha) \cdot v_k$
- $L_{y_k}(x) = f(y_k) = \nabla f(y_k)^\top (x - y_k) + \frac{\mu}{2}\|x - y_k\|_2^2$
- $L_{k+1}(x) = \psi_{k+1} + \frac{\mu}{2}\|x - v_{k+1}\|_2^2 = \beta L_k(x) + (1-\beta) L_{y_{k+1}}(x)$
- $x_{k+1} = y_k - \frac{1}{L}\nabla f(y_k)$

$$\kappa = L/\mu,\ \alpha = \frac{\sqrt{\kappa}}{\sqrt{\kappa}+1},\ \beta = 1 - \kappa^{-1/2}$$

Pick α and β so extra Terms Cancel

$$f(x_{k+1}) - \psi_{k+1} \le \beta\left[\left(\frac{\alpha(1-\beta)}{1-\alpha}\right) f(x_k) - \psi_k\right] + \beta\left[1 - \left(\frac{\alpha(1-\beta)}{1-\alpha}\right)\right] f(y_k) + \frac{1}{2}\left[\frac{(1-\beta)^2}{\mu} - \frac{1}{L}\right]\|\nabla f(y_k)\|_2^2$$

Choice of β

- $\frac{(1-\beta)^2}{\mu} - \frac{1}{L} = 0$
- $\Leftrightarrow (1-\beta)^2 = \kappa^{-1}$
- $\Leftrightarrow \beta = 1 - \kappa^{-1/2}$

Choice of α

- $\frac{\alpha(1-\beta)}{1-\alpha} = 1 \Leftrightarrow \frac{\alpha}{1-\alpha} = \frac{1}{1-\beta} = \sqrt{\kappa}$
- $\Leftrightarrow \alpha = \frac{\sqrt{\kappa}}{\sqrt{\kappa}+1}$

Proof Plan

- $y_k = \alpha \cdot x_k + (1-\alpha) \cdot v_k$
- $L_{y_k}(x) = f(y_k) = \nabla f(y_k)^\top (x - y_k) + \frac{\mu}{2}\|x - y_k\|_2^2$
- $L_{k+1}(x) = \psi_{k+1} + \frac{\mu}{2}\|x - v_{k+1}\|_2^2 = \beta L_k(x) + (1-\beta) L_{y_{k+1}}(x)$
- $x_{k+1} = y_k - \frac{1}{L}\nabla f(y_k)$

Theorem: $L_k(x) \geq f(x)$ for all $k \geq 0$ and $x \in \mathbb{R}^n$ and if $\kappa = \frac{L}{\mu}$, $\alpha = \frac{\sqrt{\kappa}}{\sqrt{\kappa}+1}$, and $\beta = 1 - \frac{1}{\sqrt{\kappa}}$, then

$$f(x_{k+1}) - \psi_{k+1} \leq \left(1 - \frac{1}{\sqrt{\kappa}}\right)[f(x_k) - \psi_k]$$

and $\sim\sqrt{\kappa}$ iterations suffices

Plan (*since $L_k(x) \geq f(x)$ fact is immediate*)

- Upper bound: $f(x_{k+1}) \leq f(y_k) - \frac{1}{2L}\|\nabla f(y_k)\|_2^2$
- Lower bound: $\psi_{k+1} \geq \beta\psi_k + (1-\beta)\left[f(y_k) - \frac{1-\beta}{2\mu}\|\nabla f(y_k)\|_2^2 + \beta\nabla f(y_k)^\top (v_k - y_k)\right]$
- Choice of y_k: $\nabla f(y_k)^\top (v_k - y_k) \geq \left(\frac{\alpha}{1-\alpha}\right)[f(y_k) - f(x_k)]$
- Pick α and β so everything cancels *(more algebra)*

Accelerated Gradient Descent (AGD)

- Initial $x_0 \in \mathbb{R}^n$, $L_0(x) = \psi_0 + \frac{\mu}{2}\|x - v_0\|_2^2$ s.t. $f(x) \geq L_0(x)$ for all x
- Repeat for $k = 0,1,2, \ldots$
  - $y_k = \alpha \cdot x_k + (1-\alpha) \cdot v_k$ where $\alpha \in [0,1]$
  - $L_{y_k}(x) = f(y_k) + \nabla f(y_k)^\top (x - y_k) + \frac{\mu}{2}\|x - y_k\|_2^2$
  - $L_{k+1}(x) = \psi_{k+1} + \frac{\mu}{2}\|x - v_{k+1}\|_2^2 = \beta L_k(x) + (1-\beta) L_{y_{k+1}}(x)$ where $\beta \in [0,1]$
  - $x_{k+1} = y_k - \frac{1}{L}\nabla f(y_k)$

$$v_{k+1} = \beta v_k + (1-\beta)\left[y_k - \frac{1}{\mu}\nabla f(y_k\right.$$

Theorem: $L_k(x) \geq f(x)$ for all $k \geq 0$ and $x \in \mathbb{R}^n$. If $\kappa = \frac{L}{\mu}$, $\alpha = \frac{\sqrt{\kappa}}{\sqrt{\kappa}+1}$, and $\beta = 1 - \kappa^{-1/2}$, then

$$f(x_{k+1}) - \psi_{k+1} \leq \left(1 - \frac{1}{\sqrt{\kappa}}\right)[f(x_k) - \psi_k]$$

and $\sim\sqrt{\kappa}$ iterations suffices

How obtain L_0?

Initial Lower Bound?

- **Goal**: $L_0(x) = \psi_0 + \frac{\mu}{2}\|x - v_0\|_2^2$ s.t. $f(x) \geq L_0(x)$
- **Idea**: $L_{x_0}(x) + f(x_0) = \nabla f(x_0)^\top (x - x_0) + \frac{\mu}{2}\|x - x_0\|_2^2$
  - $L_{x_0} = \psi_0 + \frac{\mu}{2}\|x - v_0\|_2^2$
  - $\psi_0 = f(x_0) - \frac{1}{2\mu}\|\nabla f(x_0)\|_2^2$
  - $v_0 = x_0 - \frac{1}{\mu}\nabla f(x_0)$
- **<u>One gradient evaluation!</u>**

A Proof!!

- For initial $x_0 \in \mathbb{R}^n$ compute $v_0 = x_0 - \frac{1}{\mu}\nabla f(x_0)$
- Repeat for $k = 0,1,2, \ldots$
  - $y_k = \alpha \cdot x_k + (1-\alpha) \cdot v_k$ where $\alpha = \frac{\sqrt{\kappa}}{\sqrt{\kappa}+1}$ and $\kappa = \frac{L}{\mu}$
  - $v_{k+1} = \beta v_k + (1-\beta)\left[y_k - \frac{1}{\mu}\nabla f(y_k)\right]$ *where* $\beta = 1 - \frac{1}{\sqrt{\kappa}}$
  - $x_{k+1} = y_k - \frac{1}{L}\nabla f(y_k)$

- **Theorem**: $f(x_{k+1}) - \psi_{k+1} \leq \left(1 - \frac{1}{\sqrt{\kappa}}\right)[f(x_k) - \psi_k]$ for all $k \geq 0$ where each $\psi_k \geq f(x_*)$ and $\psi_0 = f(x_0) - \frac{1}{2\mu}\|\nabla f(x_0)\|_2^2$

- **Corollary**: Can compute ϵ-optimal point in $O(\sqrt{\kappa}\log(\kappa[f(x_0) - f_*]/\epsilon))$ queries !!!
- **Proof**: $\|\nabla f(x_0)\|_2^2 \leq 2L[f(x_0) - f_*]$ and $f(x_k) - f_* \leq \left(1 - \kappa^{-1/2}\right)^k \cdot 2\kappa[f(x_0) - f_*]$

Plan for Today

- Accelerated Gradient Descent (AGD)

- Approximately optimal AGD for smooth strongly convex functions.

- Non-strongly convex
- Optimal complexity
- Momentum

Generalizations and applications

A Proof!!

How to improve?

- For initial $x_0 \in \mathbb{R}^n$ compute $v_0 = x_0 - \frac{1}{\mu}\nabla f(x_0)$
- Repeat for $k = 0,1,2, \ldots$
  - $y_k = \alpha \cdot x_k + (1-\alpha) \cdot v_k$ where $\alpha = \frac{\sqrt{\kappa}}{\sqrt{\kappa}+1}$ and $\kappa = \frac{L}{\mu}$
  - $v_{k+1} = \beta v_k + (1-\beta)\left[y_k - \frac{1}{\mu}\nabla f(y_k)\right]$ *where* $\beta = 1 - \frac{1}{\sqrt{\kappa}}$
  - $x_{k+1} = y_k - \frac{1}{L}\nabla f(y_k)$

- **Theorem**: $f(x_{k+1}) - \psi_{k+1} \leq \left(1 - \frac{1}{\sqrt{\kappa}}\right)[f(x_k) - \psi_k]$ for all $k \geq 0$ where each $\psi_k \geq f(x_*)$ and $\psi_0 = f(x_0) - \frac{1}{2\mu}\|\nabla f(x_0)\|_2^2$

- **Corollary**: Can compute ϵ-optimal point in $O(\sqrt{\kappa}\log(\kappa[f(x_0) - f_*]/\epsilon))$ queries !!!
- **Proof**: $\|\nabla f(x_0)\|_2^2 \leq 2L[f(x_0) - f_*]$ and $f(x_k) - f_* \leq \left(1 - \kappa^{-1/2}\right)^k \cdot 2\kappa[f(x_0) - f_*]$

Improved Potential Function

- For initial $x_0 \in \mathbb{R}^n$ let $v_0 = x_0$
- Repeat for $k = 0,1,2, \ldots$
  - $y_k = \alpha \cdot x_k + (1-\alpha) \cdot v_k$ where $\alpha = \frac{\sqrt{\kappa}}{\sqrt{\kappa}+1}$ and $\kappa = \frac{L}{\mu}$
  - $v_{k+1} = \beta v_k + (1-\beta)\left[y_k - \frac{1}{\mu}\nabla f(y_k)\right]$ *where* $\beta = 1 - \frac{1}{\sqrt{\kappa}}$
  - $x_{k+1} = y_k - \frac{1}{L}\nabla f(y_k)$

- **Theorem**: $p_k = f(x_k) - f_* + \frac{\mu}{2}\|v_k - x_*\|_2^2$ satisfies $p_{k+1} \leq \left(1 - \kappa^{-1/2}\right)p_k$ for all $k \geq 0$

- **Corollary**: Can compute ϵ-optimal point in $O(\sqrt{\kappa}\log([f(x_0) - f_*]/\epsilon))$ queries !!!
- **Proof**: $\frac{\mu}{2}\|x_0 - x_*\|_2^2 \leq f(x_0) - f_*$
- **Proof**: $f(x_k) - f_* \leq p_k \leq \left(1 - \kappa^{-\frac{1}{2}}\right)^k p_0 \leq \left(1 - \kappa^{-\frac{1}{2}}\right)^k \cdot 2[f(x_0) - f_*]$

Momentum?

$\kappa = \frac{L}{\mu}$, $\alpha = \frac{\sqrt{\kappa}}{\sqrt{\kappa}+1}$, and $\beta = 1 - \frac{1}{\sqrt{\kappa}}$

Algorithm 1 *(initial $x_0 \in \mathbb{R}^n$)*

- Let $v_0 = x_0$
- Repeat for $k = 0,1,2, \ldots$
  - $y_k = \alpha \cdot x_k + (1-\alpha) \cdot v_k$
  - $v_{k+1} = \beta v_k + (1-\beta)\left[y_k - \frac{1}{\mu}\nabla f(y_k)\right]$
  - $x_{k+1} = y_k - \frac{1}{L}\nabla f(y_k)$

Algorithm 2 *(initial $x_0 \in \mathbb{R}^n$)*

- Let $x_1 = x_0 - \frac{1}{L}\nabla f(x_0)$
- Repeat for $k = 1,2, \ldots$
  - $y_k = x_k + \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)[x_k - x_{k-1}]$
  - $x_{k+1} = y_k - \frac{1}{L}\nabla f(y_k)$

These algorithm are equivalent!

The x_k are identical in each algorithm.

What if not strongly convex?

Problem

$$\min_{x \in \mathbb{R}^n} f(x)$$

Idea

- $\min_x g(x) = f(x) + \frac{\lambda}{2}\|x - x_0\|_2^2$
- $g(x)$ is λ-strongly convex
- Can compute x_T an $\frac{\epsilon}{2}$-optimal point in $O\left(\sqrt{\frac{L+\lambda}{\lambda}} \log\left(\frac{g(x_0)-g_*}{\epsilon}\right)\right)$ steps
- $f(x) \leq g(x)$ so $g_* \geq f_*$
- $g(x_0) - g_* = f(x_0) - g_* \leq f(x_0) - f_* \leq \frac{L}{2}\|x_0 - x_*\|_2^2$
- $f(x_T) \leq g(x_T) \leq g_* + \epsilon \leq f(x_*) + \frac{\lambda}{2}\|x_0 - x_*\|_2^2 + \epsilon$
- If $\lambda = \frac{\epsilon}{\|x_0 - x_*\|_2^2}$ have ϵ optimal point in $O\left(\sqrt{\frac{L\|x_0 - x_*\|_2^2}{\epsilon}} \log\left(\frac{L\|x_0 - x_*\|_2^2}{\epsilon}\right)\right)$ queries

Can remove the log factor by both a better reduction and a more direct algorithm (see notes)

Plan for Today

- Accelerated Gradient Descent (AGD)

- Approximately optimal AGD for smooth strongly convex functions.

- Non-strongly convex
- Optimal complexity
- Momentum

Thursday

Generalizations and applications