Convex Optimization

(EE227A: UC Berkeley)

Lecture 13
(Gradient methods)

05 March, 2013

Suvrit Sra

Organizational

- HW2 deadline now **7th March, 2013**
- Project guidelines now on course website
- Email me to schedule meeting if you need
- Midterm on: **19th March, 2013** (in class or take home?)

Recap

- ♠ $x^{k+1} = P_{\mathcal{X}}(x^k - \alpha_k g^k)$
- ♠ Different choices of α_k (const, diminishing, Polyak)
- ♠ Can be slow; tuning α_k not so nice
- ♠ **How to decide when to stop?**
- ♠ Some other subgradient methods

Differentiable optimization

$$\min \quad f_0(x) \qquad \text{s.t. } f_i(x) \leq 0, \; i = 1, \ldots, m$$

Differentiable optimization

$$\min \quad f_0(x) \qquad \text{s.t. } f_i(x) \le 0,\ i = 1, \ldots, m$$

KKT Necessary conditions

$$\begin{aligned}
f_i(x^*) &\le 0, \quad i = 1, \ldots, m && \text{(primal feasibility)} \\
\lambda_i^* &\ge 0, \quad i = 1, \ldots, m && \text{(dual feasibility)} \\
\lambda_i^* f_i(x^*) &= 0, \quad i = 1, \ldots, m && \text{(compl. slackness)} \\
\nabla_x \mathcal{L}(x, \lambda^*)|_{x=x^*} &= 0 && \text{(Lagrangian stationarity)}
\end{aligned}$$

Differentiable optimization

$$\min \quad f_0(x) \qquad \text{s.t. } f_i(x) \le 0,\ i = 1, \ldots, m$$

KKT Necessary conditions

$$\begin{aligned}
f_i(x^*) &\le 0, \quad i = 1, \ldots, m & \text{(primal feasibility)}\\
\lambda_i^* &\ge 0, \quad i = 1, \ldots, m & \text{(dual feasibility)}\\
\lambda_i^* f_i(x^*) &= 0, \quad i = 1, \ldots, m & \text{(compl. slackness)}\\
\nabla_x \mathcal{L}(x, \lambda^*)|_{x=x^*} &= 0 & \text{(Lagrangian stationarity)}
\end{aligned}$$

Could try to solve these directly!

Nonlinear equations; sometimes solvable directly

Differentiable optimization

$$\min \quad f_0(x) \qquad \text{s.t. } f_i(x) \le 0,\ i = 1, \ldots, m$$

KKT Necessary conditions

$$\begin{aligned}
f_i(x^*) &\le 0, \quad i = 1, \ldots, m && \text{(primal feasibility)} \\
\lambda_i^* &\ge 0, \quad i = 1, \ldots, m && \text{(dual feasibility)} \\
\lambda_i^* f_i(x^*) &= 0, \quad i = 1, \ldots, m && \text{(compl. slackness)} \\
\nabla_x \mathcal{L}(x, \lambda^*)|_{x=x^*} &= 0 && \text{(Lagrangian stationarity)}
\end{aligned}$$

Could try to solve these directly!
Nonlinear equations; sometimes solvable directly

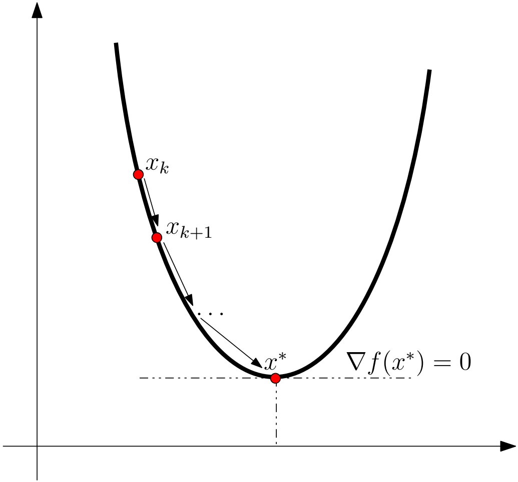
Usually quite hard; so we'll discuss iterative methods

Descent methods

$$\boxed{\min_x \quad f(x)}$$

Descent methods

$$\min_x \quad f(x)$$



Gradient methods

Unconstrained optimization

$$\min \quad f(x) \qquad x \in \mathbb{R}^n.$$

Gradient methods

Unconstrained optimization

$$\min \quad f(x) \qquad x \in \mathbb{R}^n.$$

- Suppose we have a vector $x \in \mathbb{R}^n$ for which $\nabla f(x) \neq 0$

Gradient methods

Unconstrained optimization

$$\min \quad f(x) \qquad x \in \mathbb{R}^n.$$

- Suppose we have a vector $x \in \mathbb{R}^n$ for which $\nabla f(x) \neq 0$
- Consider the ray: $x(\alpha) = x - \alpha \nabla f(x)$ for $\alpha \geq 0$

Gradient methods

Unconstrained optimization

$$\min \quad f(x) \qquad x \in \mathbb{R}^n.$$

- Suppose we have a vector $x \in \mathbb{R}^n$ for which $\nabla f(x) \neq 0$
- Consider the ray: $x(\alpha) = x - \alpha \nabla f(x)$ for $\alpha \geq 0$
- As before, make first-order Taylor expansion around x

$$f(x(\alpha)) \quad = \quad f(x) + \langle \nabla f(x), x(\alpha) - x \rangle + o(\|x(\alpha) - x\|_2)$$

Unconstrained optimization

$$\min \quad f(x) \qquad x \in \mathbb{R}^n.$$

- Suppose we have a vector $x \in \mathbb{R}^n$ for which $\nabla f(x) \neq 0$
- Consider the ray: $x(\alpha) = x - \alpha \nabla f(x)$ for $\alpha \geq 0$
- As before, make first-order Taylor expansion around x

$$\begin{aligned} f(x(\alpha)) &= f(x) + \langle \nabla f(x), x(\alpha) - x \rangle + o(\|x(\alpha) - x\|_2) \\ &= f(x) - \alpha \|\nabla f(x)\|_2^2 + o(\alpha \|\nabla f(x)\|_2) \end{aligned}$$

Gradient methods

Unconstrained optimization

$$\min \quad f(x) \qquad x \in \mathbb{R}^n.$$

- Suppose we have a vector $x \in \mathbb{R}^n$ for which $\nabla f(x) \neq 0$
- Consider the ray: $x(\alpha) = x - \alpha \nabla f(x)$ for $\alpha \geq 0$
- As before, make first-order Taylor expansion around x

$$\begin{aligned} f(x(\alpha)) &= f(x) + \langle \nabla f(x), x(\alpha) - x \rangle + o(\|x(\alpha) - x\|_2) \\ &= f(x) - \alpha \|\nabla f(x)\|_2^2 + o(\alpha \|\nabla f(x)\|_2) \\ &= f(x) - \alpha \|\nabla f(x)\|_2^2 + o(\alpha) \end{aligned}$$

Gradient methods

Unconstrained optimization

$$\min \quad f(x) \qquad x \in \mathbb{R}^n.$$

- Suppose we have a vector $x \in \mathbb{R}^n$ for which $\nabla f(x) \neq 0$
- Consider the ray: $x(\alpha) = x - \alpha \nabla f(x)$ for $\alpha \geq 0$
- As before, make first-order Taylor expansion around x

$$\begin{aligned} f(x(\alpha)) &= f(x) + \langle \nabla f(x), x(\alpha) - x \rangle + o(\|x(\alpha) - x\|_2) \\ &= f(x) - \alpha \|\nabla f(x)\|_2^2 + o(\alpha \|\nabla f(x)\|_2) \\ &= f(x) - \alpha \|\nabla f(x)\|_2^2 + o(\alpha) \end{aligned}$$

- For α near 0, $\alpha \|\nabla f(x)\|_2^2$ dominates $o(\alpha)$
- For positive, sufficiently small α, $f(x(\alpha))$ **smaller** than $f(x)$

Descent methods

- Carrying the idea further, consider

$$x(\alpha) = x + \alpha d,$$

where **direction** $d \in \mathbb{R}^n$ obtuse to $\nabla f(x)$, i.e.,

$$\langle \nabla f(x),\, d \rangle < 0.$$

Descent methods

- Carrying the idea further, consider

$$x(\alpha) = x + \alpha d,$$

where **direction** $d \in \mathbb{R}^n$ obtuse to $\nabla f(x)$, i.e.,

$$\langle \nabla f(x),\, d \rangle < 0.$$

- Again, we have the Taylor expansion

$$f(x(\alpha)) = f(x) + \alpha \langle \nabla f(x),\, d \rangle + o(\alpha),$$

where $\langle \nabla f(x),\, d \rangle$ dominates $o(\alpha)$ for suff. small α

Descent methods

- Carrying the idea further, consider

$$x(\alpha) = x + \alpha d,$$

where **direction** $d \in \mathbb{R}^n$ obtuse to $\nabla f(x)$, i.e.,

$$\langle \nabla f(x),\, d \rangle < 0.$$

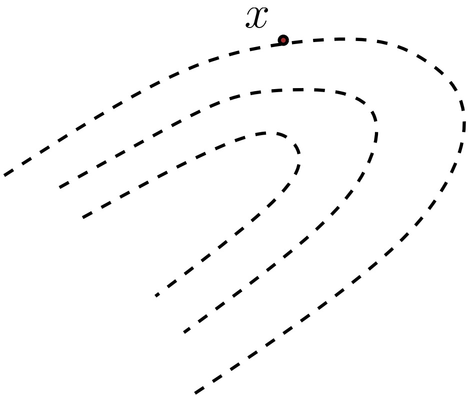
- Again, we have the Taylor expansion

$$f(x(\alpha)) = f(x) + \alpha \langle \nabla f(x),\, d \rangle + o(\alpha),$$

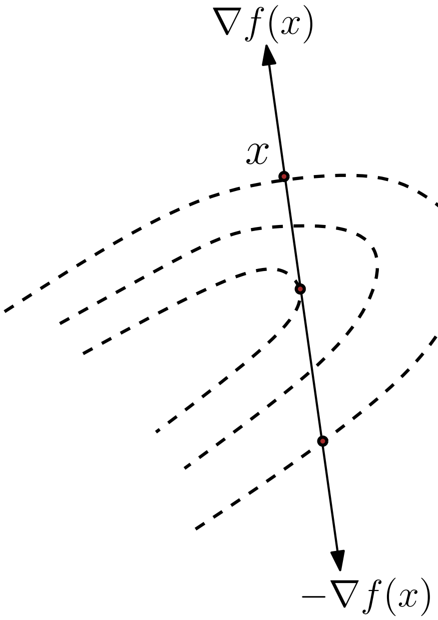
where $\langle \nabla f(x),\, d \rangle$ dominates $o(\alpha)$ for suff. small α

- Since d is obtuse to $\nabla f(x)$, this implies $f(x(\alpha)) < f(x)$

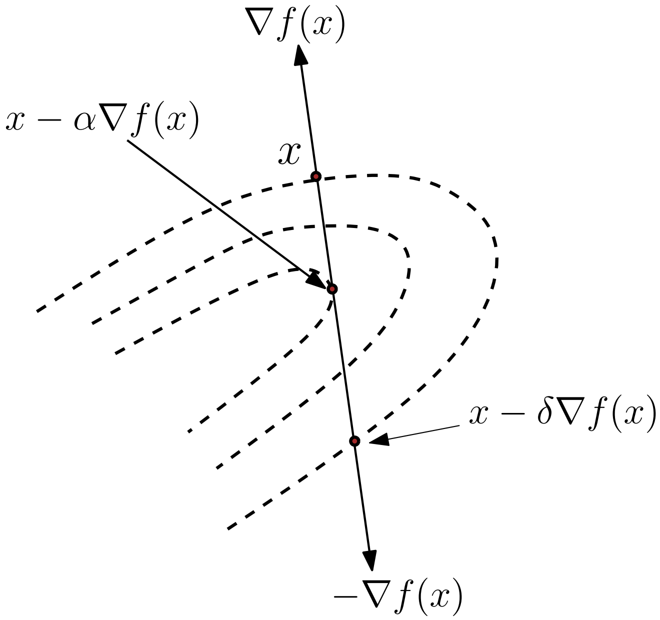
Descent methods



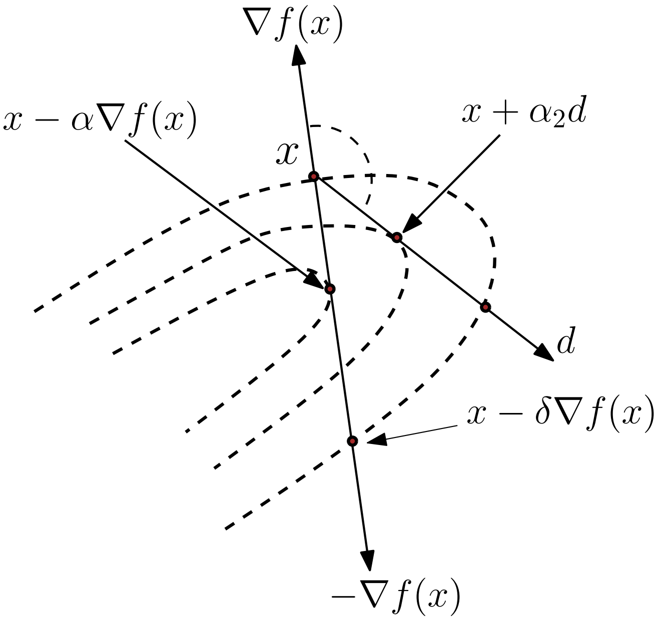
Descent methods



Descent methods



Descent methods



Algorithm

1. Start with some guess x^0;
2. For each $k = 0, 1, \ldots$
   - $x^{k+1} \leftarrow x^k + \alpha_k d^k$
   - Check when to stop (e.g., if $\nabla f(x^{k+1}) = 0$)

Gradient methods

$$\boxed{x^{k+1} = x^k + \alpha_k d^k, \quad k = 0, 1, \ldots}$$

Gradient methods

$$\boxed{x^{k+1} = x^k + \alpha_k d^k, \quad k = 0, 1, \ldots}$$

- **stepsize** $\alpha_k \geq 0$, usually ensures $f(x^{k+1}) < f(x^k)$

Gradient methods

$$\boxed{x^{k+1} = x^k + \alpha_k d^k, \quad k = 0, 1, \ldots}$$

- **stepsize** $\alpha_k \geq 0$, usually ensures $f(x^{k+1}) < f(x^k)$
- **Descent direction** d^k satisfies

$$\langle \nabla f(x^k),\, d^k \rangle < 0$$

Gradient methods

$$x^{k+1} = x^k + \alpha_k d^k, \quad k = 0, 1, \ldots$$

- **stepsize** $\alpha_k \geq 0$, usually ensures $f(x^{k+1}) < f(x^k)$
- **Descent direction** d^k satisfies

$$\langle \nabla f(x^k), d^k \rangle < 0$$

Numerous ways to select α_k and d^k

Gradient methods

$$\boxed{x^{k+1} = x^k + \alpha_k d^k, \quad k = 0, 1, \ldots}$$

- **stepsize** $\alpha_k \geq 0$, usually ensures $f(x^{k+1}) < f(x^k)$
- **Descent direction** d^k satisfies

$$\langle \nabla f(x^k), d^k \rangle < 0$$

$$\boxed{\text{Numerous ways to select } \alpha_k \text{ and } d^k}$$

Usually methods **seek** **monotonic descent**

$$f(x^{k+1}) < f(x^k)$$

Generic matlab code

```
function [x,f] = gradientDescent(x0)

  fx  = @(x)  objfn(x);    % handle to f(x)
  gfx = @(x) grad(x);      % handle to nabla f(x)

  x=x0;                    % input starting point
  maxiter = 100;           % tunable parameter

  for k=1:maxiter          % or other criterion
     g = gfx(x);           % compute gradient at x
     al = stepSize(x);     % compute a stepsize
     x = x - al*g;         % perform update
     fprintf('Iter: %d\t Obj: %d\n', fx(x));
  end
end
```

Gradient methods – direction

$$x^{k+1} = x^k + \alpha_k d^k, \quad k = 0, 1, \ldots$$

- Different choices of direction d^k

- **Scaled gradient:** $d^k = -D^k \nabla f(x^k)$, $D^k \succ 0$
- **Newton's method:** $(D^k = [\nabla^2 f(x^k)]^{-1})$
- **Quasi-Newton:** $D^k \approx [\nabla^2 f(x^k)]^{-1}$
- **Steepest descent:** $D^k = \boldsymbol{I}$
- **Diagonally scaled:** D^k diagonal with $D^k_{ii} \approx \left(\frac{\partial^2 f(x^k)}{(\partial x_i)^2}\right)^{-1}$
- **Discretized Newton:** $D^k = [H(x^k)]^{-1}$, H via finite-diff.

Gradient methods – direction

$$x^{k+1} = x^k + \alpha_k d^k, \quad k = 0, 1, \ldots$$

- Different choices of direction d^k

- **Scaled gradient:** $d^k = -D^k \nabla f(x^k)$, $D^k \succ 0$
- **Newton's method:** $(D^k = [\nabla^2 f(x^k)]^{-1})$
- **Quasi-Newton:** $D^k \approx [\nabla^2 f(x^k)]^{-1}$
- **Steepest descent:** $D^k = \boldsymbol{I}$
- **Diagonally scaled:** D^k diagonal with $D^k_{ii} \approx \left(\frac{\partial^2 f(x^k)}{(\partial x_i)^2}\right)^{-1}$
- **Discretized Newton:** $D^k = [H(x^k)]^{-1}$, H via finite-diff.
- $\ldots$

Exercise: Verify that $\langle \nabla f(x^k), d^k \rangle < 0$ for above choices

Gradient methods – stepsize

- **Exact:** $\alpha_k := \operatorname*{argmin}_{\alpha \geq 0} f(x^k + \alpha d^k)$

Gradient methods – stepsize

- **Exact:** $\alpha_k := \underset{\alpha \geq 0}{\operatorname{argmin}} f(x^k + \alpha d^k)$
- **Limited min:** $\alpha_k = \underset{0 \leq \alpha \leq s}{\operatorname{argmin}} f(x^k + \alpha d^k)$

Gradient methods – stepsize

- **Exact:** $\alpha_k := \underset{\alpha \geq 0}{\operatorname{argmin}} f(x^k + \alpha d^k)$
- **Limited min:** $\alpha_k = \underset{0 \leq \alpha \leq s}{\operatorname{argmin}} f(x^k + \alpha d^k)$
- **Armijo-rule**. Given **fixed** scalars, s, β, σ with $0 < \beta < 1$ and $0 < \sigma < 1$ (chosen experimentally). Set

$$\alpha_k = \beta^{m_k} s,$$

where we **try** $\beta^m s$ for $m = 0, 1, \ldots$ until **sufficient descent**

$$f(x^k) - f(x + \beta^m s d^k) \geq -\sigma \beta^m s \langle \nabla f(x^k),\, d^k \rangle$$

Gradient methods – stepsize

- **Exact:** $\alpha_k := \underset{\alpha \geq 0}{\operatorname{argmin}} f(x^k + \alpha d^k)$
- **Limited min:** $\alpha_k = \underset{0 \leq \alpha \leq s}{\operatorname{argmin}} f(x^k + \alpha d^k)$
- **Armijo-rule**. Given **fixed** scalars, s, β, σ with $0 < \beta < 1$ and $0 < \sigma < 1$ (chosen experimentally). Set

$$\alpha_k = \beta^{m_k} s,$$

where we **try** $\beta^m s$ for $m = 0, 1, \ldots$ until **sufficient descent**

$$f(x^k) - f(x + \beta^m s d^k) \geq -\sigma \beta^m s \langle \nabla f(x^k),\, d^k \rangle$$

If $\langle \nabla f(x^k),\, d^k \rangle < 0$, stepsize guaranteed to exist

Gradient methods – stepsize

- **Exact:** $\alpha_k := \underset{\alpha \geq 0}{\operatorname{argmin}} f(x^k + \alpha d^k)$
- **Limited min:** $\alpha_k = \underset{0 \leq \alpha \leq s}{\operatorname{argmin}} f(x^k + \alpha d^k)$
- **Armijo-rule**. Given **fixed** scalars, s, β, σ with $0 < \beta < 1$ and $0 < \sigma < 1$ (chosen experimentally). Set

$$\alpha_k = \beta^{m_k} s,$$

where we **try** $\beta^m s$ for $m = 0, 1, \ldots$ until **sufficient descent**

$$f(x^k) - f(x + \beta^m s d^k) \geq -\sigma \beta^m s \langle \nabla f(x^k),\, d^k \rangle$$

If $\langle \nabla f(x^k),\, d^k \rangle < 0$, stepsize guaranteed to exist
Usually, σ small $\in [10^{-5}, 0.1]$, while β from $1/2$ to $1/10$ depending on how confident we are about initial stepsize s.

Gradient methods – stepsize

- **Constant:** $\alpha_k = 1/L$ (for suitable value of L)
- **Diminishing:** $\alpha_k \to 0$ but $\sum_k \alpha_k = \infty$.

Gradient methods – stepsize

- **Constant:** $\alpha_k = 1/L$ (for suitable value of L)
- **Diminishing:** $\alpha_k \to 0$ but $\sum_k \alpha_k = \infty$.
  Latter condition ensures that $\{x^k\}$ does not converge to nonstationary points.

Gradient methods – stepsize

- **Constant:** $\alpha_k = 1/L$ (for suitable value of L)
- **Diminishing:** $\alpha_k \to 0$ but $\sum_k \alpha_k = \infty$.
  Latter condition ensures that $\{x^k\}$ does not converge to nonstationary points.
  Say, $x^k \to \bar{x}$; then for sufficiently large m and n, $(m > n)$

$$x^m \approx x^n \approx \bar{x},$$

Gradient methods – stepsize

- **Constant:** $\alpha_k = 1/L$ (for suitable value of L)
- **Diminishing:** $\alpha_k \to 0$ but $\sum_k \alpha_k = \infty$.
  Latter condition ensures that $\{x^k\}$ does not converge to nonstationary points.
  Say, $x^k \to \bar{x}$; then for sufficiently large m and n, $(m > n)$

$$x^m \approx x^n \approx \bar{x},\ x^m \approx x^n - \left(\sum_{k=n}^{m-1} \alpha_k\right) \nabla f(\bar{x}).$$

Gradient methods – stepsize

- **Constant:** $\alpha_k = 1/L$ (for suitable value of L)
- **Diminishing:** $\alpha_k \to 0$ but $\sum_k \alpha_k = \infty$.
  Latter condition ensures that $\{x^k\}$ does not converge to nonstationary points.
  Say, $x^k \to \bar{x}$; then for sufficiently large m and n, $(m > n)$

  $$x^m \approx x^n \approx \bar{x}, \; x^m \approx x^n - \left(\sum_{k=n}^{m-1} \alpha_k\right) \nabla f(\bar{x}).$$

  The sum can be made arbitrarily large, contradicting nonstationarity of $\bar{x}$

Gradient methods – nonmonotonic steps*

- Stepsize computation can be expensive
- Convergence analysis depends on monotonic descent

Gradient methods – nonmonotonic steps*

- Stepsize computation can be expensive
- Convergence analysis depends on monotonic descent
- Give up search for stepsizes
- Use closed-form formulae for stepsizes
- Don't insist on monotonic descent?
- (e.g., diminishing stepsizes do not enforce monotonic descent)

Gradient methods – nonmonotonic steps*

- Stepsize computation can be expensive
- Convergence analysis depends on monotonic descent
- Give up search for stepsizes
- Use closed-form formulae for stepsizes
- Don't insist on monotonic descent?
- (e.g., diminishing stepsizes do not enforce monotonic descent)

Barzilai & Borwein stepsizes

$$x^{k+1} = x^k - \alpha^k \nabla f(x^k), \quad k = 0, 1, \ldots$$

Gradient methods – nonmonotonic steps*

- Stepsize computation can be expensive
- Convergence analysis depends on monotonic descent
- Give up search for stepsizes
- Use closed-form formulae for stepsizes
- Don't insist on monotonic descent?
- (e.g., diminishing stepsizes do not enforce monotonic descent)

Barzilai & Borwein stepsizes

$$x^{k+1} = x^k - \alpha^k \nabla f(x^k), \quad k = 0, 1, \ldots$$

$$\alpha_k = \frac{\langle u^k,\, v^k \rangle}{\|v^k\|^2}, \qquad \alpha_k = \frac{\|u^k\|^2}{\langle u^k,\, v^k \rangle}$$

$$u^k = x^k - x^{k-1}, \quad v^k = \nabla f(x^k) - \nabla f(x^{k-1})$$

Barzilai-Borwein steps – remarks

Intriguing behavior:

- ♠ Akin to simultaneous descent-direction $\times$ step
- ♠ Result in *non-monotonic* descent
- ♠ Work quite well empirically
- ♠ Good for **large-scale problems**
- ♠ Difficult convergence analysis
- ♠ Often supplemented with nonmonotonic line-search

Convergence theory

Gradient descent – convergence

$$\boxed{x^{k+1} = x^k - \alpha_k \nabla f(x^k), \quad k = 0, 1, \ldots}$$

Gradient descent – convergence

$$\boxed{x^{k+1} = x^k - \alpha_k \nabla f(x^k), \quad k = 0, 1, \ldots}$$

Convergence

Theorem $\|\nabla f(x^k)\|_2 \to 0$ as $k \to \infty$

Gradient descent – convergence

$$\boxed{x^{k+1} = x^k - \alpha_k \nabla f(x^k), \quad k = 0, 1, \ldots}$$

Convergence

Theorem $\|\nabla f(x^k)\|_2 \to 0$ as $k \to \infty$

Convergence rate with constant stepsize

Theorem Let $f \in C_L^1$ and $\{x^k\}$ be sequence generated as above, with $\alpha_k = 1/L$. Then, $f(x^{T+1}) - f(x^*) = O(1/T)$.

Gradient descent – convergence

$$\boxed{x^{k+1} = x^k - \alpha_k \nabla f(x^k), \quad k = 0, 1, \ldots}$$

Convergence

Theorem $\|\nabla f(x^k)\|_2 \to 0$ as $k \to \infty$

Convergence rate with constant stepsize

Theorem Let $f \in C_L^1$ and $\{x^k\}$ be sequence generated as above, with $\alpha_k = 1/L$. Then, $f(x^{T+1}) - f(x^*) = O(1/T)$.

Proof plan:

- Show that $f(x^{k+1}) < f(x^k)$ (for suitable L)
- Measure progress via $\|x^k - x^*\|_2^2$ as before
- Sum up bounds, induct to obtain rate

Gradient descent – convergence

Assumption: Lipschitz continuous gradient; denoted $f \in C_L^1$

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2$$

Gradient descent – convergence

Assumption: Lipschitz continuous gradient; denoted $f \in C^1_L$

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2$$

- ♣ Gradient vectors of closeby points are close to each other
- ♣ Objective function has "bounded curvature"
- ♣ Speed at which gradient varies is bounded

Gradient descent – convergence

Assumption: Lipschitz continuous gradient; denoted $f \in C_L^1$

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2$$

- ♣ Gradient vectors of closeby points are close to each other
- ♣ Objective function has “bounded curvature”
- ♣ Speed at which gradient varies is bounded

Lemma (Descent). Let $f \in C_L^1$. Then,

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

Gradient descent – convergence

Assumption: Lipschitz continuous gradient; denoted $f \in C_L^1$

$$\|\nabla f(x) - \nabla f(y)\|_2 \le L\|x - y\|_2$$

- ♣ Gradient vectors of closeby points are close to each other
- ♣ Objective function has "bounded curvature"
- ♣ Speed at which gradient varies is bounded

Lemma (Descent). Let $f \in C_L^1$. Then,

$$f(x) \le f(y) + \langle \nabla f(y), x - y\rangle + \frac{L}{2}\|x - y\|_2^2$$

For convex f, compare with

$$f(x) \ge f(y) + \langle \nabla f(y), x - y\rangle.$$

Descent lemma

Proof. Since $f \in C_L^1$, by Taylor's theorem, for the vector $z_t = y + t(x - y)$ we have

$$f(x) = f(y) + \int_0^1 \langle \nabla f(z_t), x - y \rangle dt.$$

Descent lemma

Proof. Since $f \in C_L^1$, by Taylor's theorem, for the vector $z_t = y + t(x - y)$ we have

$$f(x) = f(y) + \int_0^1 \langle \nabla f(z_t), x - y \rangle dt.$$

Add and subtract $\langle \nabla f(y), x - y \rangle$ on rhs we have

$$f(x) - f(y) - \langle \nabla f(y), x - y \rangle \quad = \quad \int_0^1 \langle \nabla f(z_t) - \nabla f(y), x - y \rangle dt$$

Descent lemma

Proof. Since $f \in C_L^1$, by Taylor's theorem, for the vector $z_t = y + t(x-y)$ we have

$$f(x) = f(y) + \int_0^1 \langle \nabla f(z_t),\, x-y \rangle dt.$$

Add and subtract $\langle \nabla f(y),\, x-y \rangle$ on rhs we have

$$\begin{aligned} f(x) - f(y) - \langle \nabla f(y),\, x-y \rangle &= \int_0^1 \langle \nabla f(z_t) - \nabla f(y),\, x-y \rangle dt \\ |f(x) - f(y) - \langle \nabla f(y),\, x-y \rangle| &\leq \left| \int_0^1 \langle \nabla f(z_t) - \nabla f(y),\, x-y \rangle dt \right| \end{aligned}$$

Descent lemma

Proof. Since $f \in C_L^1$, by Taylor's theorem, for the vector $z_t = y + t(x - y)$ we have

$$f(x) = f(y) + \int_0^1 \langle \nabla f(z_t), x - y \rangle dt.$$

Add and subtract $\langle \nabla f(y), x - y \rangle$ on rhs we have

$$\begin{aligned}
f(x) - f(y) - \langle \nabla f(y), x - y \rangle &= \int_0^1 \langle \nabla f(z_t) - \nabla f(y), x - y \rangle dt \\
|f(x) - f(y) - \langle \nabla f(y), x - y \rangle| &\leq \left| \int_0^1 \langle \nabla f(z_t) - \nabla f(y), x - y \rangle dt \right| \\
&\leq \int_0^1 |\langle \nabla f(z_t) - \nabla f(y), x - y \rangle| dt
\end{aligned}$$

Descent lemma

Proof. Since $f \in C_L^1$, by Taylor's theorem, for the vector $z_t = y + t(x-y)$ we have

$$f(x) = f(y) + \int_0^1 \langle \nabla f(z_t),\, x-y \rangle dt.$$

Add and subtract $\langle \nabla f(y),\, x-y \rangle$ on rhs we have

$$\begin{aligned}
f(x) - f(y) - \langle \nabla f(y),\, x-y \rangle &= \int_0^1 \langle \nabla f(z_t) - \nabla f(y),\, x-y \rangle dt \\
|f(x) - f(y) - \langle \nabla f(y),\, x-y \rangle| &\leq \left| \int_0^1 \langle \nabla f(z_t) - \nabla f(y),\, x-y \rangle dt \right| \\
&\leq \int_0^1 |\langle \nabla f(z_t) - \nabla f(y),\, x-y \rangle| dt \\
&\leq \int_0^1 \|\nabla f(z_t) - \nabla f(y)\|_2 x - y dt
\end{aligned}$$

Descent lemma

Proof. Since $f \in C_L^1$, by Taylor's theorem, for the vector $z_t = y + t(x-y)$ we have

$$f(x) = f(y) + \int_0^1 \langle \nabla f(z_t),\, x - y\rangle dt.$$

Add and subtract $\langle \nabla f(y),\, x-y\rangle$ on rhs we have

$$\begin{aligned}
f(x) - f(y) - \langle \nabla f(y),\, x-y\rangle &= \int_0^1 \langle \nabla f(z_t) - \nabla f(y),\, x-y\rangle dt \\
|f(x) - f(y) - \langle \nabla f(y),\, x-y\rangle| &\leq \left| \int_0^1 \langle \nabla f(z_t) - \nabla f(y),\, x-y\rangle dt \right| \\
&\leq \int_0^1 |\langle \nabla f(z_t) - \nabla f(y),\, x-y\rangle| dt \\
&\leq \int_0^1 \|\nabla f(z_t) - \nabla f(y)\|_2 x - y dt \\
&\leq L \int_0^1 t\|x-y\|_2^2 dt
\end{aligned}$$

Descent lemma

Proof. Since $f \in C_L^1$, by Taylor's theorem, for the vector $z_t = y + t(x-y)$ we have

$$f(x) = f(y) + \int_0^1 \langle \nabla f(z_t), x-y\rangle dt.$$

Add and subtract $\langle \nabla f(y), x-y\rangle$ on rhs we have

$$\begin{aligned} f(x) - f(y) - \langle \nabla f(y), x-y\rangle &= \int_0^1 \langle \nabla f(z_t) - \nabla f(y), x-y\rangle dt \\ |f(x) - f(y) - \langle \nabla f(y), x-y\rangle| &\leq \left|\int_0^1 \langle \nabla f(z_t) - \nabla f(y), x-y\rangle dt\right| \\ &\leq \int_0^1 |\langle \nabla f(z_t) - \nabla f(y), x-y\rangle| dt \\ &\leq \int_0^1 \|\nabla f(z_t) - \nabla f(y)\|_2 x - y dt \\ &\leq L \int_0^1 t\|x-y\|_2^2 dt \\ &= \frac{L}{2}\|x-y\|_2^2. \end{aligned}$$

Bounds $f(x)$ above and below with quadratic functions

Descent lemma – corollaries

Coroll. 1 If $f \in C_L^1$, and $0 < \alpha_k < 2/L$, then $f(x^{k+1}) < f(x^k)$

$$f(x^{k+1}) \quad \leq \quad f(x^k) + \langle \nabla f(x^k), x^{k+1} - x^k \rangle + \tfrac{L}{2} \|x^{k+1} - x^k\|_2$$

Descent lemma – corollaries

Coroll. 1 If $f \in C_L^1$, and $0 < \alpha_k < 2/L$, then $f(x^{k+1}) < f(x^k)$

$$
\begin{aligned}
f(x^{k+1}) &\leq f(x^k) + \langle \nabla f(x^k), x^{k+1} - x^k \rangle + \tfrac{L}{2}\|x^{k+1} - x^k\|_2 \\
&= f(x^k) - \alpha_k \|\nabla f(x^k)\|_2^2 + \tfrac{\alpha_k^2 L}{2}\|\nabla f(x^k)\|_2^2
\end{aligned}
$$

Descent lemma – corollaries

Coroll. 1 If $f \in C^1_L$, and $0 < \alpha_k < 2/L$, then $f(x^{k+1}) < f(x^k)$

$$
\begin{aligned}
f(x^{k+1}) &\leq f(x^k) + \langle \nabla f(x^k),\, x^{k+1} - x^k \rangle + \tfrac{L}{2}\|x^{k+1} - x^k\|_2 \\
&= f(x^k) - \alpha_k \|\nabla f(x^k)\|_2^2 + \tfrac{\alpha_k^2 L}{2}\|\nabla f(x^k)\|_2^2 \\
&= f(x^k) - \alpha_k\left(1 - \tfrac{\alpha_k}{2} L\right)\|\nabla f(x^k)\|_2^2
\end{aligned}
$$

Descent lemma – corollaries

Coroll. 1 If $f \in C^1_L$, and $0 < \alpha_k < 2/L$, then $f(x^{k+1}) < f(x^k)$

$$
\begin{aligned}
f(x^{k+1}) &\leq f(x^k) + \langle \nabla f(x^k), x^{k+1} - x^k \rangle + \tfrac{L}{2}\|x^{k+1} - x^k\|_2 \\
&= f(x^k) - \alpha_k \|\nabla f(x^k)\|_2^2 + \tfrac{\alpha_k^2 L}{2}\|\nabla f(x^k)\|_2^2 \\
&= f(x^k) - \alpha_k(1 - \tfrac{\alpha_k}{2}L)\|\nabla f(x^k)\|_2^2
\end{aligned}
$$

Thus, if $\alpha_k < 2/L$ we have descent.

Descent lemma – corollaries

Coroll. 1 If $f \in C^1_L$, and $0 < \alpha_k < 2/L$, then $f(x^{k+1}) < f(x^k)$

$$\begin{aligned} f(x^{k+1}) &\leq f(x^k) + \langle \nabla f(x^k),\, x^{k+1} - x^k \rangle + \tfrac{L}{2}\|x^{k+1} - x^k\|_2 \\ &= f(x^k) - \alpha_k \|\nabla f(x^k)\|_2^2 + \tfrac{\alpha_k^2 L}{2}\|\nabla f(x^k)\|_2^2 \\ &= f(x^k) - \alpha_k(1 - \tfrac{\alpha_k}{2}L)\|\nabla f(x^k)\|_2^2 \end{aligned}$$

Thus, if $\alpha_k < 2/L$ we have descent. Minimize over α_k to get best bound: this yields $\alpha_k = 1/L$—**we'll use this stepsize**

Convergence

- We showed that

$$f(x^k) - f(x^{k+1}) \geq \frac{c}{L}\|\nabla f(x^k)\|_2^2,$$

($c = 1/2$ for $\alpha_k = 1/L$; c has diff. value for other stepsize rules)

Convergence

- We showed that

$$f(x^k) - f(x^{k+1}) \geq \tfrac{c}{L}\|\nabla f(x^k)\|_2^2,$$

($c = 1/2$ for $\alpha_k = 1/L$; c has diff. value for other stepsize rules)

- Sum up above inequalities for $k = 0, 1, \ldots, T$ to obtain

$$\frac{c}{L}\sum_{k=0}^{T}\|\nabla f(x^k)\|_2^2 \quad \leq \quad f(x^0) - f(x^{T+1})$$

Convergence

- We showed that

$$f(x^k) - f(x^{k+1}) \geq \tfrac{c}{L}\|\nabla f(x^k)\|_2^2,$$

($c = 1/2$ for $\alpha_k = 1/L$; c has diff. value for other stepsize rules)

- Sum up above inequalities for $k = 0, 1, \ldots, T$ to obtain

$$\frac{c}{L}\sum_{k=0}^{T}\|\nabla f(x^k)\|_2^2 \quad \leq \quad f(x^0) - f(x^{T+1}) \leq f(x^0) - f^*$$

Convergence

- We showed that

$$f(x^k) - f(x^{k+1}) \geq \tfrac{c}{L}\|\nabla f(x^k)\|_2^2,$$

$(c = 1/2$ for $\alpha_k = 1/L$; c has diff. value for other stepsize rules)

- Sum up above inequalities for $k = 0, 1, \ldots, T$ to obtain

$$\frac{c}{L}\sum_{k=0}^{T}\|\nabla f(x^k)\|_2^2 \quad \leq \quad f(x^0) - f(x^{T+1}) \leq f(x^0) - f^*$$

- We assume $f^* > -\infty$, so rhs is some fixed positive constant

Convergence

- We showed that

$$f(x^k) - f(x^{k+1}) \geq \tfrac{c}{L}\|\nabla f(x^k)\|_2^2,$$

($c = 1/2$ for $\alpha_k = 1/L$; c has diff. value for other stepsize rules)

- Sum up above inequalities for $k = 0, 1, \ldots, T$ to obtain

$$\frac{c}{L}\sum_{k=0}^{T}\|\nabla f(x^k)\|_2^2 \quad \leq \quad f(x^0) - f(x^{T+1}) \leq f(x^0) - f^*$$

- We assume $f^* > -\infty$, so rhs is some fixed positive constant
- Thus, as $k \to \infty$, lhs must converge; thus

$$\|\nabla f(x^k)\|_2 \to 0 \quad \text{as} \quad k \to \infty.$$

Convergence

- We showed that

$$f(x^k) - f(x^{k+1}) \geq \tfrac{c}{L}\|\nabla f(x^k)\|_2^2,$$

($c = 1/2$ for $\alpha_k = 1/L$; c has diff. value for other stepsize rules)

- Sum up above inequalities for $k = 0, 1, \ldots, T$ to obtain

$$\frac{c}{L}\sum_{k=0}^{T}\|\nabla f(x^k)\|_2^2 \quad \leq \quad f(x^0) - f(x^{T+1}) \leq f(x^0) - f^*$$

- We assume $f^* > -\infty$, so rhs is some fixed positive constant
- Thus, as $k \to \infty$, lhs must converge; thus

$$\|\nabla f(x^k)\|_2 \to 0 \quad \text{as} \quad k \to \infty.$$

- Notice, we **did not require** f to be convex . . .

Descent lemma – another corollary

Corollary 2 If f is a **convex** function $\in C_L^1$, then

$$\frac{1}{L}\|\nabla f(x) - \nabla f(y)\|_2^2 \leq \langle \nabla f(x) - \nabla f(y), x - y \rangle,$$

Exercise: Prove this corollary.

Convergence rate – convex f

- $\star$ Let $\alpha_k = 1/L$
- $\star$ Shorthand notation $g^k = \nabla f(x^k)$, $g^* = \nabla f(x^*)$
- $\star$ Let $r_k := \|x^k - x^*\|_2$ (distance to optimum)

Convergence rate – convex f

- ⋆ Let $\alpha_k = 1/L$
- ⋆ Shorthand notation $g^k = \nabla f(x^k)$, $g^* = \nabla f(x^*)$
- ⋆ Let $r_k := \|x^k - x^*\|_2$ (distance to optimum)

Lemma Distance to min shrinks monotonically; $r_{k+1} \leq r_k$

Convergence rate – convex f

- ⋆ Let $\alpha_k = 1/L$
- ⋆ Shorthand notation $g^k = \nabla f(x^k)$, $g^* = \nabla f(x^*)$
- ⋆ Let $r_k := \|x^k - x^*\|_2$ (distance to optimum)

Lemma Distance to min shrinks monotonically; $r_{k+1} \leq r_k$

Proof. Descent lemma implies that: $f(x^{k+1}) \leq f(x^k) - \frac{1}{2L}\|g^k\|_2^2$

Convergence rate – convex f

- ⋆ Let $\alpha_k = 1/L$
- ⋆ Shorthand notation $g^k = \nabla f(x^k)$, $g^* = \nabla f(x^*)$
- ⋆ Let $r_k := \|x^k - x^*\|_2$ (distance to optimum)

Lemma Distance to min shrinks monotonically; $r_{k+1} \leq r_k$

Proof. Descent lemma implies that: $f(x^{k+1}) \leq f(x^k) - \frac{1}{2L}\|g^k\|_2^2$

Consider, $r_{k+1}^2 = \|x^{k+1} - x^*\|_2^2 = \|x^k - x^* - \alpha_k g^k\|_2^2$.

Convergence rate – convex f

- $\star$ Let $\alpha_k = 1/L$
- $\star$ Shorthand notation $g^k = \nabla f(x^k)$, $g^* = \nabla f(x^*)$
- $\star$ Let $r_k := \|x^k - x^*\|_2$ (distance to optimum)

Lemma Distance to min shrinks monotonically; $r_{k+1} \leq r_k$

Proof. Descent lemma implies that: $f(x^{k+1}) \leq f(x^k) - \frac{1}{2L}\|g^k\|_2^2$

Consider, $r_{k+1}^2 = \|x^{k+1} - x^*\|_2^2 = \|x^k - x^* - \alpha_k g^k\|_2^2$.

$$r_{k+1}^2 \;=\; r_k^2 + \alpha_k^2\|g^k\|_2^2 - 2\alpha_k\langle g^k,\, x^k - x^*\rangle$$

Convergence rate – convex f

- $\star$ Let $\alpha_k = 1/L$
- $\star$ Shorthand notation $g^k = \nabla f(x^k)$, $g^* = \nabla f(x^*)$
- $\star$ Let $r_k := \|x^k - x^*\|_2$ (distance to optimum)

Lemma Distance to min shrinks monotonically; $r_{k+1} \leq r_k$

Proof. Descent lemma implies that: $f(x^{k+1}) \leq f(x^k) - \frac{1}{2L}\|g^k\|_2^2$

Consider, $r_{k+1}^2 = \|x^{k+1} - x^*\|_2^2 = \|x^k - x^* - \alpha_k g^k\|_2^2$.

$$\begin{aligned} r_{k+1}^2 &= r_k^2 + \alpha_k^2\|g^k\|_2^2 - 2\alpha_k\langle g^k,\, x^k - x^*\rangle \\ &= r_k^2 + \alpha_k^2\|g^k\|_2^2 - 2\alpha_k\langle g^k - g^*,\, x^k - x^*\rangle \quad \text{as } g^* = 0 \end{aligned}$$

Convergence rate – convex f

- $\star$ Let $\alpha_k = 1/L$
- $\star$ Shorthand notation $g^k = \nabla f(x^k)$, $g^* = \nabla f(x^*)$
- $\star$ Let $r_k := \|x^k - x^*\|_2$ (distance to optimum)

Lemma Distance to min shrinks monotonically; $r_{k+1} \leq r_k$

Proof. Descent lemma implies that: $f(x^{k+1}) \leq f(x^k) - \frac{1}{2L}\|g^k\|_2^2$

Consider, $r_{k+1}^2 = \|x^{k+1} - x^*\|_2^2 = \|x^k - x^* - \alpha_k g^k\|_2^2$.

$$\begin{aligned} r_{k+1}^2 &= r_k^2 + \alpha_k^2\|g^k\|_2^2 - 2\alpha_k\langle g^k, x^k - x^*\rangle \\ &= r_k^2 + \alpha_k^2\|g^k\|_2^2 - 2\alpha_k\langle g^k - g^*, x^k - x^*\rangle \quad \text{as } g^* = 0 \\ &\leq r_k^2 + \alpha_k^2\|g^k\|_2^2 - \frac{2\alpha_k}{L}\|g^k - g^*\|_2^2 \qquad \text{(Coroll. 2)} \end{aligned}$$

Convergence rate – convex f

- ⋆ Let $\alpha_k = 1/L$
- ⋆ Shorthand notation $g^k = \nabla f(x^k)$, $g^* = \nabla f(x^*)$
- ⋆ Let $r_k := \|x^k - x^*\|_2$ (distance to optimum)

Lemma Distance to min shrinks monotonically; $r_{k+1} \leq r_k$

Proof. Descent lemma implies that: $f(x^{k+1}) \leq f(x^k) - \frac{1}{2L}\|g^k\|_2^2$

Consider, $r_{k+1}^2 = \|x^{k+1} - x^*\|_2^2 = \|x^k - x^* - \alpha_k g^k\|_2^2$.

$$\begin{aligned}
r_{k+1}^2 &= r_k^2 + \alpha_k^2\|g^k\|_2^2 - 2\alpha_k\langle g^k, x^k - x^*\rangle \\
&= r_k^2 + \alpha_k^2\|g^k\|_2^2 - 2\alpha_k\langle g^k - g^*, x^k - x^*\rangle \quad \text{as } g^* = 0 \\
&\leq r_k^2 + \alpha_k^2\|g^k\|_2^2 - \tfrac{2\alpha_k}{L}\|g^k - g^*\|_2^2 \qquad \text{(Coroll. 2)} \\
&= r_k^2 - \alpha_k(\tfrac{2}{L} - \alpha_k)\|g^k\|_2^2.
\end{aligned}$$

Convergence rate – convex f

- ⋆ Let $\alpha_k = 1/L$
- ⋆ Shorthand notation $g^k = \nabla f(x^k)$, $g^* = \nabla f(x^*)$
- ⋆ Let $r_k := \|x^k - x^*\|_2$ (distance to optimum)

Lemma Distance to min shrinks monotonically; $r_{k+1} \leq r_k$

Proof. Descent lemma implies that: $f(x^{k+1}) \leq f(x^k) - \frac{1}{2L}\|g^k\|_2^2$

Consider, $r_{k+1}^2 = \|x^{k+1} - x^*\|_2^2 = \|x^k - x^* - \alpha_k g^k\|_2^2$.

$$\begin{aligned} r_{k+1}^2 &= r_k^2 + \alpha_k^2\|g^k\|_2^2 - 2\alpha_k\langle g^k,\, x^k - x^*\rangle \\ &= r_k^2 + \alpha_k^2\|g^k\|_2^2 - 2\alpha_k\langle g^k - g^*,\, x^k - x^*\rangle \quad \text{as } g^* = 0 \\ &\leq r_k^2 + \alpha_k^2\|g^k\|_2^2 - \tfrac{2\alpha_k}{L}\|g^k - g^*\|_2^2 \qquad \text{(Coroll. 2)} \\ &= r_k^2 - \alpha_k(\tfrac{2}{L} - \alpha_k)\|g^k\|_2^2. \end{aligned}$$

Since $\alpha_k < 2/L$, it follows that $r_{k+1} \leq r_k$

Convergence rate

Lemma Let $\Delta_k := f(x^k) - f(x^*)$. Then, $\Delta_{k+1} \leq \Delta_k(1-\beta)$

Convergence rate

Lemma Let $\Delta_k := f(x^k) - f(x^*)$. Then, $\Delta_{k+1} \leq \Delta_k(1-\beta)$

$$f(x^k) - f(x^*) = \Delta_k \overset{\text{cvx } f}{\leq} \langle g^k, x^k - x^* \rangle$$

Convergence rate

Lemma Let $\Delta_k := f(x^k) - f(x^*)$. Then, $\Delta_{k+1} \leq \Delta_k(1-\beta)$

$$f(x^k) - f(x^*) = \Delta_k \overset{\text{cvx } f}{\leq} \langle g^k,\, x^k - x^* \rangle \quad \overset{\text{CS}}{\leq} \quad \|g^k\|_2 \underbrace{\|x^k - x^*\|_2}_{r_k}.$$

Convergence rate

Lemma Let $\Delta_k := f(x^k) - f(x^*)$. Then, $\Delta_{k+1} \leq \Delta_k(1-\beta)$

$$f(x^k) - f(x^*) = \Delta_k \overset{\text{cvx } f}{\leq} \langle g^k, x^k - x^* \rangle \overset{\text{CS}}{\leq} \|g^k\|_2 \underbrace{\|x^k - x^*\|_2}_{r_k}.$$

That is, $\|g^k\|_2 \geq \Delta_k / r_k$.

Convergence rate

Lemma Let $\Delta_k := f(x^k) - f(x^*)$. Then, $\Delta_{k+1} \le \Delta_k(1-\beta)$

$$f(x^k) - f(x^*) = \Delta_k \overset{\text{cvx } f}{\le} \langle g^k, x^k - x^* \rangle \overset{\text{CS}}{\le} \|g^k\|_2 \underbrace{\|x^k - x^*\|_2}_{r_k}.$$

That is, $\|g^k\|_2 \ge \Delta_k / r_k$. In particular, since $r_k \le r_0$, we have

$$\|g^k\|_2 \ge \frac{\Delta_k}{r_0}.$$

Convergence rate

Lemma Let $\Delta_k := f(x^k) - f(x^*)$. Then, $\Delta_{k+1} \leq \Delta_k(1-\beta)$

$$f(x^k) - f(x^*) = \Delta_k \overset{\text{cvx } f}{\leq} \langle g^k, x^k - x^* \rangle \overset{\text{CS}}{\leq} \|g^k\|_2 \underbrace{\|x^k - x^*\|_2}_{r_k}.$$

That is, $\|g^k\|_2 \geq \Delta_k / r_k$. In particular, since $r_k \leq r_0$, we have

$$\|g^k\|_2 \geq \frac{\Delta_k}{r_0}.$$

Now we have a bound on the gradient norm...

Convergence rate

Recall $f(x^{k+1}) \le f(x^k) - \frac{1}{2L}\|g^k\|_2^2$; subtracting f^* from both sides

$$\Delta_{k+1} \le \Delta_k - \frac{\Delta_k^2}{2Lr_0^2} = \Delta_k\big(1 - \tfrac{\Delta_k}{2Lr_0^2}\big)$$

Convergence rate

Recall $f(x^{k+1}) \leq f(x^k) - \frac{1}{2L}\|g^k\|_2^2$; subtracting f^* from both sides

$$\Delta_{k+1} \leq \Delta_k - \frac{\Delta_k^2}{2Lr_0^2} = \Delta_k\big(1 - \tfrac{\Delta_k}{2Lr_0^2}\big) = \Delta_k(1-\beta).$$

Convergence rate

Recall $f(x^{k+1}) \leq f(x^k) - \frac{1}{2L}\|g^k\|_2^2$; subtracting f^* from both sides

$$\Delta_{k+1} \leq \Delta_k - \frac{\Delta_k^2}{2Lr_0^2} = \Delta_k\big(1 - \tfrac{\Delta_k}{2Lr_0^2}\big) = \Delta_k(1-\beta).$$

$$\boxed{\text{But we want to bound: } f(x^{T+1}) - f(x^*)}$$

Convergence rate

Recall $f(x^{k+1}) \le f(x^k) - \frac{1}{2L}\|g^k\|_2^2$; subtracting f^* from both sides

$$\Delta_{k+1} \le \Delta_k - \frac{\Delta_k^2}{2Lr_0^2} = \Delta_k\big(1 - \tfrac{\Delta_k}{2Lr_0^2}\big) = \Delta_k(1-\beta).$$

$$\boxed{\text{But we want to bound: } f(x^{T+1}) - f(x^*)}$$

$$\implies \quad \frac{1}{\Delta_{k+1}} \ge \frac{1}{\Delta_k}(1+\beta) = \frac{1}{\Delta_k} + \frac{1}{2Lr_0^2}$$

Convergence rate

Recall $f(x^{k+1}) \leq f(x^k) - \frac{1}{2L}\|g^k\|_2^2$; subtracting f^* from both sides

$$\Delta_{k+1} \leq \Delta_k - \frac{\Delta_k^2}{2Lr_0^2} = \Delta_k\big(1 - \tfrac{\Delta_k}{2Lr_0^2}\big) = \Delta_k(1-\beta).$$

$$\boxed{\text{But we want to bound: } f(x^{T+1}) - f(x^*)}$$

$$\Longrightarrow \quad \frac{1}{\Delta_{k+1}} \geq \frac{1}{\Delta_k}(1+\beta) = \frac{1}{\Delta_k} + \frac{1}{2Lr_0^2}$$

- Sum both sides over $k = 0, \ldots, T$ to obtain

$$\frac{1}{\Delta_{T+1}} \geq \frac{1}{\Delta_0} + \frac{T+1}{2Lr_0^2}$$

Convergence rate

- Sum both sides over $k = 0, \dots, T$ to obtain

$$\frac{1}{\Delta_{T+1}} \geq \frac{1}{\Delta_0} + \frac{T+1}{2Lr_0^2}$$

Convergence rate

- Sum both sides over $k = 0, \ldots, T$ to obtain

$$\frac{1}{\Delta_{T+1}} \geq \frac{1}{\Delta_0} + \frac{T+1}{2Lr_0^2}$$

- Rearrange to conclude

$$f(x^T) - f^* \leq \frac{2L\Delta_0 r_0^2}{2Lr_0^2 + T\Delta_0}$$

Convergence rate

- Sum both sides over $k = 0, \ldots, T$ to obtain

$$\frac{1}{\Delta_{T+1}} \geq \frac{1}{\Delta_0} + \frac{T+1}{2Lr_0^2}$$

- Rearrange to conclude

$$f(x^T) - f^* \leq \frac{2L\Delta_0 r_0^2}{2Lr_0^2 + T\Delta_0}$$

- Use descent lemma to bound $\Delta_0 \leq (L/2)\|x^0 - x^*\|_2^2$; simplify

$$f(x^T) - f(x^*) \leq \frac{2L\Delta_0\|x^0 - x^*\|_2^2}{T+4} = O(1/T).$$

Exercise: Prove above simplification.

Gradient descent – faster rate

Assumption: **Strong convexity**; denote $f \in S^1_{L,\mu}$

$$f(x) \geq f(y) + \langle \nabla f(y), x - y \rangle + \frac{\mu}{2}\|x - y\|_2^2$$

- Setting $\alpha_k = 2/(\mu + L)$ yields linear rate $(\mu > 0)$

Strongly convex case

Thm 2. Suppose $f \in S^1_{L,\mu}$. Then, for any $x, y \in \mathbb{R}^n$

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \frac{\mu L}{\mu + L} \|x - y\|_2^2 + \frac{1}{\mu + L} \|\nabla f(x) - \nabla f(y)\|_2^2$$

Strongly convex case

Thm 2. Suppose $f \in S^1_{L,\mu}$. Then, for any $x, y \in \mathbb{R}^n$

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \frac{\mu L}{\mu + L} \|x - y\|_2^2 + \frac{1}{\mu + L} \|\nabla f(x) - \nabla f(y)\|_2^2$$

- Consider the **convex** function $\phi(x) = f(x) - \frac{\mu}{2}\|x\|_2^2$

Strongly convex case

Thm 2. Suppose $f \in S^1_{L,\mu}$. Then, for any $x, y \in \mathbb{R}^n$

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \frac{\mu L}{\mu + L} \|x - y\|_2^2 + \frac{1}{\mu + L} \|\nabla f(x) - \nabla f(y)\|_2^2$$

- Consider the **convex** function $\phi(x) = f(x) - \frac{\mu}{2}\|x\|_2^2$
- $\nabla \phi(x) = \nabla f(x) - \mu x$

Strongly convex case

Thm 2. Suppose $f \in S^1_{L,\mu}$. Then, for any $x, y \in \mathbb{R}^n$

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \frac{\mu L}{\mu + L}\|x - y\|_2^2 + \frac{1}{\mu + L}\|\nabla f(x) - \nabla f(y)\|_2^2$$

- Consider the **convex** function $\phi(x) = f(x) - \frac{\mu}{2}\|x\|_2^2$
- $\nabla \phi(x) = \nabla f(x) - \mu x$
- If $\mu = L$, then easily true (due to strong convexity and Coroll. 2)

Strongly convex case

Thm 2. Suppose $f \in S^1_{L,\mu}$. Then, for any $x, y \in \mathbb{R}^n$

$$\langle \nabla f(x) - \nabla f(y),\, x - y \rangle \geq \frac{\mu L}{\mu + L}\|x - y\|_2^2 + \frac{1}{\mu + L}\|\nabla f(x) - \nabla f(y)\|_2^2$$

- Consider the **convex** function $\phi(x) = f(x) - \frac{\mu}{2}\|x\|_2^2$
- $\nabla \phi(x) = \nabla f(x) - \mu x$
- If $\mu = L$, then easily true (due to strong convexity and Coroll. 2)
- If $\mu < L$, then $\phi \in C^1_{L-\mu}$; now invoke Coroll. 2

$$\langle \nabla \phi(x) - \nabla \phi(y),\, x - y \rangle \geq \frac{1}{L - \mu}\|\nabla \phi(x) - \nabla \phi(y)\|_2$$

Strongly convex – rate

Theorem. If $f \in S^1_{L,\mu}$, $0 < \alpha < 2/(L+\mu)$, then the gradient method generates a sequence $\{x^k\}$ that satisfies

$$\|x^k - x^*\|_2^2 \leq \left(1 - \frac{2\alpha\mu L}{\mu + L}\right)^k \|x^0 - x^*\|_2.$$

Moreover, if $\alpha = 2/(L+\mu)$ then

$$f(x^k) - f^* \leq \frac{L}{2}\left(\frac{\kappa - 1}{\kappa + 1}\right)^{2k} \|x^0 - x^*\|_2^2,$$

where $\kappa = L/\mu$ is the **condition number**.

Strongly convex – rate

- As before, let $r_k = \|x^k - x^*\|_2$, and consider

Strongly convex – rate

- As before, let $r_k = \|x^k - x^*\|_2$, and consider

$$r_{k+1}^2 \quad = \quad \|x^k - x^* - \alpha \nabla f(x^k)\|_2^2$$

Strongly convex – rate

- As before, let $r_k = \|x^k - x^*\|_2$, and consider

$$\begin{aligned} r_{k+1}^2 &= \|x^k - x^* - \alpha \nabla f(x^k)\|_2^2 \\ &= r_k^2 - 2\alpha\langle \nabla f(x^k),\, x^k - x^*\rangle + \alpha^2 \|\nabla f(x^k)\|_2^2 \end{aligned}$$

Strongly convex – rate

- As before, let $r_k = \|x^k - x^*\|_2$, and consider

$$
\begin{aligned}
r_{k+1}^2 &= \|x^k - x^* - \alpha\nabla f(x^k)\|_2^2 \\
&= r_k^2 - 2\alpha\langle\nabla f(x^k),\, x^k - x^*\rangle + \alpha^2\|\nabla f(x^k)\|_2^2 \\
&\leq \left(1 - \frac{2\alpha\mu L}{\mu + L}\right) r_k^2 + \alpha\left(\alpha - \frac{2}{\mu + L}\right)\|\nabla f(x^k)\|_2^2
\end{aligned}
$$

  where we used Thm. 2 with $\nabla f(x^*) = 0$ for last inequality.

Exercise: Complete the proof using above argument.

Gradient methods – lower bounds

$$x^{k+1} = x^k - \alpha_k \nabla f(x^k)$$

Theorem Lower bound I (Nesterov) For any $x^0 \in \mathbb{R}^n$, and $1 \le T \le \frac{1}{2}(n-1)$, there is a smooth f, s.t.

$$f(x^T) - f(x^*) \ge \frac{3L\|x^0 - x^*\|_2^2}{32(T+1)^2}$$

Gradient methods – lower bounds

$$x^{k+1} = x^k - \alpha_k \nabla f(x^k)$$

Theorem Lower bound I (Nesterov) For any $x^0 \in \mathbb{R}^n$, and $1 \leq T \leq \frac{1}{2}(n-1)$, there is a smooth f, s.t.

$$f(x^T) - f(x^*) \geq \frac{3L\|x^0 - x^*\|_2^2}{32(T+1)^2}$$

Theorem Lower bound II (Nesterov). For class of smooth, strongly convex, i.e., $S^{\infty}_{L,\mu}$ ($\mu > 0$, $\kappa > 1$)

$$f(x^T) - f(x^*) \geq \frac{\mu}{2}\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{2T}\|x^0 - x^*\|_2^2.$$

We'll come back to these towards end of course

Exercise

- Let D be the $(n-1) \times n$ *differencing* matrix

$$D = \begin{pmatrix} -1 & 1 & & & & \\ & -1 & 1 & & & \\ & & & \ddots & & \\ & & & & -1 & 1 \end{pmatrix} \in \mathbb{R}^{(n-1)\times n},$$

- $f(x) = \frac{1}{2}\|D^T x - b\|_2^2 = \frac{1}{2}(\|D^T x\|_2^2 + \|b\|_2^2 - 2\langle D^T x,\, b\rangle)$
- Notice that $\nabla f(x) = D(D^T x - b)$
- Try different choices of b, and different initial vectors x_0
- **Exercise:** Experiment to see how large n must be before subgradient method starts outperforming CVX
- **Exercise:** Minimize $f(x)$ for large n; e.g., $n = 10^6$, $n = 10^7$
- **Exercise:** Repeat same exercise with constraints: $x_i \in [-1, 1]$.

References

♡ Bertsekas (1999); *"Nonlinear programming"* (1999)

♡ Nesterov (2003); *"Introductory lectures on convex optimization"*