ECE 8101: Nonconvex Optimization for Machine Learning

Lecture Note 2-3: Gradient Descent

Jia (Kevin) Liu

Associate Professor
Department of Electrical and Computer Engineering
The Ohio State University, Columbus, OH, USA

Autumn 2024

Outline

In this lecture:

- Convergence rate concept
- Gradient descent method
- Convergence performance of gradient descent
- Step size selection strategies

Iterative Algorithms for Optimization

We consider the following iterative algorithms:

$$\mathbf{x}_{k+1} = \mathbf{x}_k + s_k \mathbf{d}_k,$$

where s_k is step-size, and $\mathbf{d}_k$ is search direction depending on $(\mathbf{x}_k, \mathbf{x}_{k-1}, \ldots)$.

For now: assume f smooth, $f(\mathbf{x}_k)$ and $\nabla f(\mathbf{x}_k)$ is easy to evaluate

Complications from ML:

- Nonconvex f
- Nonsmooth f
- f not available (or too expensive to evaluate exactly)
- Only an estimate of $\nabla f(\mathbf{x}_k)$ is available
- A constraint $\mathbf{x} \in \Omega$ (usually a relatively simple Ω, e.g., ball, box, simplex...)
- Nonsmooth regularization, i.e., instead of $f(\mathbf{x})$, we want $\min f(\mathbf{x}) + \tau\psi(\mathbf{x})$

How to Evaluate the Speed of an Iterative Algorithm?

Definition 1 (Convergence rate)

A sequence $\{r_k\} \to r^*$ and $r_k \neq r^*$ for all k. The rate (or order) of convergence p is a nonnegative number satisfying

$$\limsup_{k\to\infty} \frac{\|r_{k+1} - r^*\|}{\|r_k - r^*\|^p} = \beta < \infty.$$

$\frac{\|r_{k+1} - r^*\|}{\|r_k - r^*\|} \to 1$ as $k \to \infty$.

- Sublinear: $p = 1$ and $\beta = 1$ (e.g., $O(1/k)$ rate, kind of slow but still OK)

  $\frac{c/(k+1)}{c/k} = \frac{k}{k+1} \to 1$. Desired $\varepsilon > 0$: $\frac{c}{k} \le \varepsilon \Rightarrow k \ge \frac{c}{\varepsilon} = O(\frac{1}{\varepsilon})$

- Linear or geometric: $p = 1$ and $0 < \beta < 1$ (i.e., $\|r_{k+1} - r^*\| \le \beta\|r_k - r^*\|$ for some $\beta \in (0, 1)$, or $\|r_k - r^*\| = O(\beta^k)$, which is quite fast) contraction. $\le \beta^2\|r_{k-1} - r^*\| \le \beta^k\|r_1 - r^*\| = O(\beta^k)$.

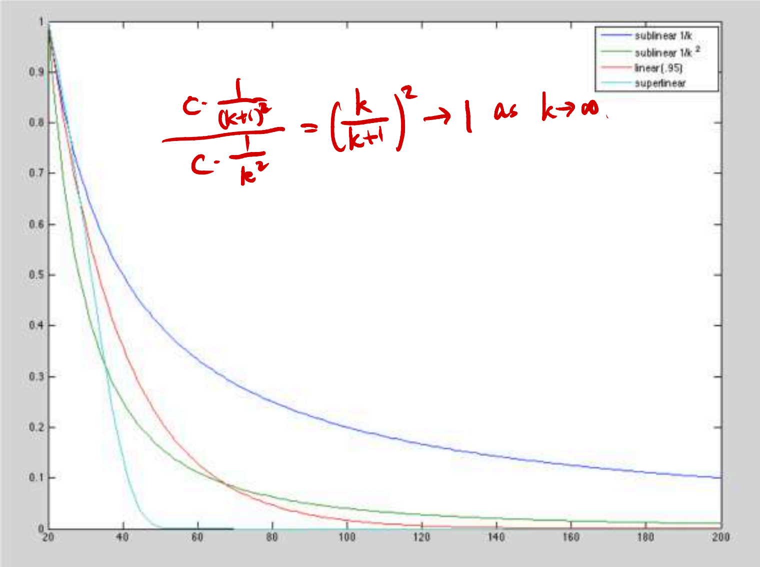
  Desired ε: $c\beta^k \le \varepsilon \Rightarrow k \ge c\log(\varepsilon^{-1})$. Need $O(\log(\varepsilon^{-1}))$ iter.

- Superlinear: $p > 1$ and $\beta < \infty$, or $p = 1$ and $\beta = 0$ (i.e., $\frac{\|r_{k+1} - r^*\|}{\|r_k - r^*\|} \to 0$, that's very fast!) Not only a contraction, but also the rate of convergence is accelerating.

- Quadratic: $p = 2$ and $\beta < \infty$ ($\|r_{k+1} - r^*\| \le \beta\|r_k - r^*\|^2$, # of correct significant digits doubles per iteration. Rarely need anything faster than this!)

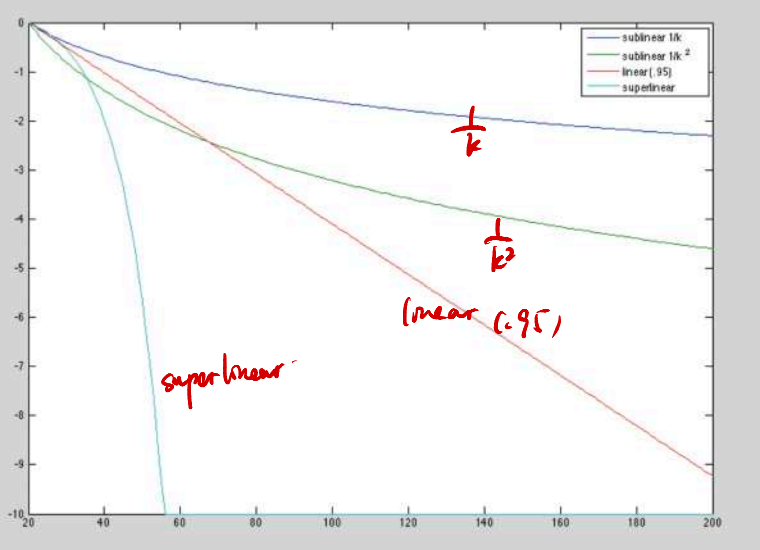
  For ε-accuracy: Need $O(\log\log(\varepsilon^{-1}))$ iter. $\leftarrow$ almost const.

Convergence Rates Comparisons



$$\frac{C \cdot \frac{1}{(k+1)^2}}{C \cdot \frac{1}{k^2}} = \left(\frac{k}{k+1}\right)^2 \to 1 \text{ as } k \to \infty.$$

Convergence Rates Comparisons: Log-Scale



Gradient Descent

Back to the unconstrained optimization problem, with f smooth and convex:

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

Denote the optimal value as $f^* = \min_{\mathbf{x}} f(\mathbf{x}^*)$ and an optimal solution as $\mathbf{x}^*$

Gradient Descent

Choose initial point $\mathbf{x}_0 \in \mathbb{R}^n$. Repeat:

$$\mathbf{x}_k = \mathbf{x}_{k-1} - s_k \nabla f(\mathbf{x}_{k-1}), \quad k = 1, 2, 3, \ldots$$

Stop if some stopping criterion is satisfied. e.g., $\|\nabla f(\underline{x}_k)\| \leq \varepsilon$.

$\|\underline{x}_{k+1} - \underline{x}_k\| \leq \varepsilon$.

stop after fixed # of iters. (finite-time conv. analysis).

Gradient Descent: Geometric Interpretation

Gradient descent is a first-order method: Consider the following quadratic Taylor approximation:

SO - approx

$$f(\mathbf{y}) \approx \underbrace{f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y}-\mathbf{x})}_{\text{FO - approx}} + \frac{1}{2}(\mathbf{y}-\mathbf{x})^\top \underbrace{\nabla^2 f(\mathbf{x})}_{\frac{1}{s}\mathbf{I}}(\mathbf{y}-\mathbf{x}) + o(\|\mathbf{y}-\mathbf{x}\|^2)$$

No, we replace Hessian $\nabla^2 f(\mathbf{x})$ by $\frac{1}{s}\mathbf{I}$ to obtain:

$$f(\mathbf{y}) \approx \underbrace{f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y}-\mathbf{x})}_{\text{FO - approx}} + \boxed{\frac{1}{2s}\|\mathbf{y}-\mathbf{x}\|^2}$$

"proximity penalty"

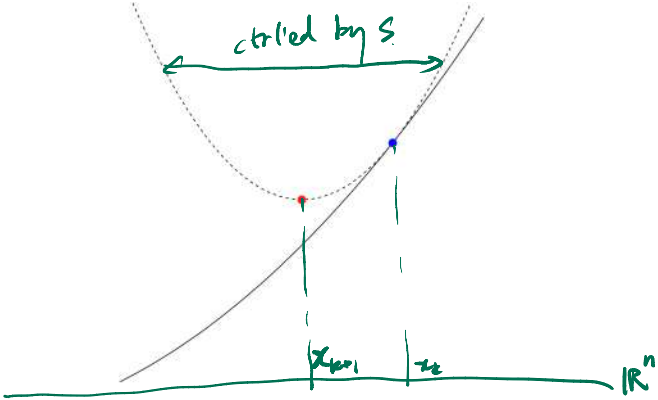
Can be viewed as a linear approximation to f, with proximity term to $\mathbf{x}$ weighted by $\frac{1}{2s}$. Choose next point $\mathbf{y} = \mathbf{x}^+$ to minimize this approximation:

$$\mathbf{x}^+ = \mathbf{x} - s\nabla f(\mathbf{x})$$

Quadratic fn of y : set grad → 0 , solve for y

$\nabla f(y) = 0 \rightarrow \nabla f(\underline{x}) + \frac{1}{s}(y - \underline{x}) \Rightarrow \boxed{y = \underline{x} - s\nabla f(\underline{x})}$

Gradient Descent: Geometric Interpretation



$$\mathbf{x}^+ = \arg\min_y f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) + \frac{1}{2s} \|\mathbf{y} - x\|_2^2$$

Questions:

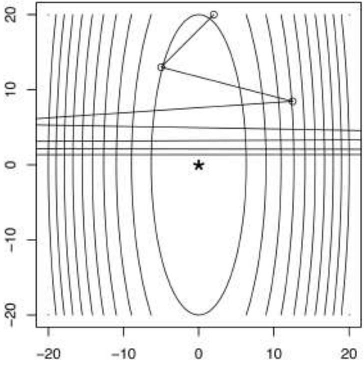
- How to choose step sizes $\{s_k\}$?
- What is the according convergence rate? Or does it depend on $\{s_k\}$?

Strategy 1: Fixed Step Size

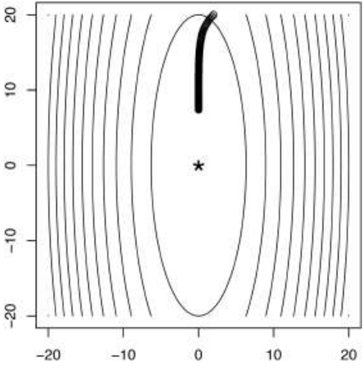
Simply set $s_k = s$ for all $k = 1, 2, 3, \ldots$.

Limitations: May diverge if s is too large, Can be slow if s is too small.

Example: Consider $f(\mathbf{x}) = (10x_1^2 + x_2^2)/2$: $\Rightarrow (x_1^*, x_2^*) = (0,0)$.



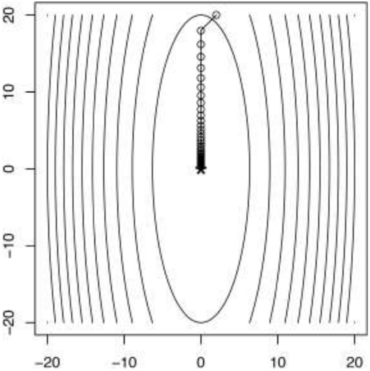
8 iterations



100 iterations

Strategy 1: Fixed Step Size

Converges nicely when s is "just right." Same example, GD after 40 iterations:



Will be clear what we mean by "just right" in convergence rate analysis later

Convergence Rate Analysis (Convex): Fixed Step Size

Assume that f is convex & differentiable, with $\mathrm{dom}(f) = \mathbb{R}^n$ and additionally

L-smooth.

$$\|\nabla f(\mathbf{y}) - \nabla f(\mathbf{x})\|_2 \leq L\|\mathbf{y} - \mathbf{x}\|_2, \quad \forall \mathbf{x}, \mathbf{y}$$

That is, ∇f is Lipschitz continuous with constant $L > 0$ (L-Lipschitz continuous)

$h: D \subseteq \mathbb{R}^n$. h is Lip. cont. : $\exists L > 0$ s.t. $\|h(y) - h(x)\| \leq L\|y - x\|$.

Theorem 1 (Optimality Gap)

If f is convex, differentiable, and L-smooth, gradient descent with fixed step size $s \leq 1/L$ satisfies

$$f(\mathbf{x}_k) - f(\mathbf{x}^*) \leq \frac{\|\mathbf{x}_0 - \mathbf{x}^*\|_2^2}{2sk},$$

i.e., gradient descent method has sublinear convergence rate $O(1/k)$.

Remark:

- To get $f(\mathbf{x}_k) - f(\mathbf{x}^*) \leq \epsilon$, it takes $O(1/\epsilon)$ iterations.

Theorem 1 (Optimality Gap)

If f is convex, differentiable, and L-smooth, gradient descent with fixed step size $s \le 1/L$ satisfies

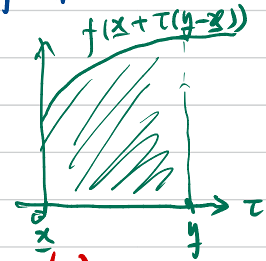
$$f(\mathbf{x}_k) - f(\mathbf{x}^*) \le \frac{\|\mathbf{x}_0 - \mathbf{x}^*\|_2^2}{2sk}, = O(\tfrac{1}{k})$$

i.e., gradient descent method has sublinear convergence rate $O(1/k)$.

Proof. Step ① Claim: If ∇f is Lipschitz, then

descent lemma

$$f(y) \le f(x) + \nabla f(x)^T (y-x) + \frac{L}{2}\|y-x\|^2, \ \forall x, y \in \mathbb{R}^n. \quad (1)$$



To show (1), we consider

$$f(y) = f(x) + \int_0^1 \underbrace{f'(x+\tau(y-x))}_{\text{dir. der.}} \, d\tau$$

$$= f(x) + \int_0^1 \nabla f(x+\tau(y-x))^T (y-x)\, d\tau \quad \text{(chain rule)}$$

$$\overset{\text{add \& sub}}{=} f(x) + \int_0^1 \left[[\nabla f(x+\tau(y-x))]^T (y-x) + \nabla f(x)^T(y-x) - \nabla f(x)^T(y-x) \right] d\tau$$

$$= f(x) + \nabla f(x)^T (y-x) + \int_0^1 [\nabla f(x+\tau(y-x)) - \nabla f(x)]^T (y-x)\, d\tau$$

By rearranging & taking abs. val. on both sides

$$|f(y) - f(x) - \nabla f(x)^T(y-x)| = \left| \int_0^1 [\nabla f(x+\tau(y-x)) - \nabla f(x)]^T (y-x)\, d\tau \right|$$

$$\le \int_0^1 \left| [\nabla f(x+\tau(y-x)) - \nabla f(x)]^T (y-x) \right| d\tau \quad \left(\text{Triangle Ineq } \|a+b\| \le \|a\| + \|b\|\right)$$

$$\le \int_0^1 \underbrace{\|\nabla f(x+\tau(y-x)) - \nabla f(x)\|}_{\text{L-Lipschitz}: \ \le L\tau\|y-x\|} \cdot \|y-x\| \, d\tau \quad \left(\text{Cauchy-Schwarz } |a^Tb| \le \|a\| \cdot \|b\|\right)$$

$$\le \int_0^1 L\tau \|y-x\|^2 d\tau = L\|y-x\|^2 \underbrace{\int_0^1 \tau \, d\tau}_{=\frac{1}{2}} = \frac{L}{2}\|y-x\|^2.$$

(1) is proved.

step ②: WTS "Descent property of GD".

$x_{k+1} = x_k - s_k \nabla f(x_k)$. Plug this into (1).

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

$$= f(x_k) - s\|\nabla f(x_k)\|^2 + \frac{Ls^2}{2}\|\nabla f(x_k)\|^2$$

General:

$$= f(x_k) - s\left(1-\frac{Ls}{2}\right)\|\nabla f(x_k)\|^2 \quad \xleftarrow[(2)]{}$$

starting pt.

Lyapunov.

$\ge 0, = 0$

$$\left.\begin{aligned} V_k - V_{k-1} &\le -\delta_{k-1} \\ &\vdots \\ V_1 - V_0 &\le -\delta_0 \end{aligned}\right\} \Rightarrow$$

$$V_k - V_0 \le -\sum_{i=0}^{k-1}\delta_i$$

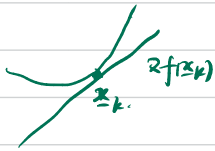
step ③: Consider $\{\|x_k - x^*\|^2\}_{k=1}^{\infty}$

check: $\|x_{k+1}-x^*\|^2 - \|x_k - x^*\|^2$

$$= \|\underline{x_k - s\nabla f(x_k) - x^*}\|^2 - \|x_k - x^*\|^2$$

$$= \cancel{\|x_k-x^*\|^2} - 2s\nabla f(x_k)^T(x_k-x^*) + s^2\|\nabla f(x_k)\|^2 - \cancel{\|x_k-x^*\|^2}$$

$$= -2s\nabla f(x_k)^T(x_k-x^*) + s^2\|\nabla f(x_k)\|^2$$



Due to convexity: $f(x^*) \ge f(x_k) + \nabla f(x_k)^T(x^*-x_k)$

$$\Rightarrow f(x_k) \le f(x^*) + f(x_k)^T(x_k - x^*). \qquad (3)$$

Plugging (3) into (2):

$$f(\underline{x}_{k+1}) \le \underbrace{f(\underline{x}^*) + \nabla f(\underline{x}_k)^T(\underline{x}_k - \underline{x}^*)}_{f(\underline{x}_k) \le} - s\left(1 - \frac{Ls}{2}\right)\|\nabla f(\underline{x}_k)\|^2. \quad (4)$$

Note that $s \in (0, \frac{1}{L}]$. Then $0 < s \le \frac{1}{L} \Rightarrow 0 < Ls \le 1$.

$$\Rightarrow -\frac{1}{2} \le -\frac{Ls}{2} < 0 \Rightarrow \frac{1}{2} \le 1 - \frac{Ls}{2} \le 1 \Rightarrow -s \le -s\left(1-\frac{Ls}{2}\right) \le -\frac{s}{2}.$$

Using this in (4):

$$f(\underline{x}_{k+1}) - f(\underline{x}^*) \le \nabla f(\underline{x}_k)^T(\underline{x}_k - \underline{x}^*) - \frac{s}{2}\|\nabla f(\underline{x}_k)\|^2$$

$$\Rightarrow -2s\nabla f(\underline{x}_k)^T(\underline{x}_k - \underline{x}^*) \le -2s\left(f(\underline{x}_{k+1}) - f(\underline{x}^*)\right) - s^2\|\nabla f(\underline{x}_k)\|^2.$$

$$\Rightarrow 2s\left(f(\underline{x}_{k+1}) - f(\underline{x}^*)\right) \le 2s\nabla f(\underline{x}_k)^T(\underline{x}_k - \underline{x}^*) - s^2\|\nabla f(\underline{x}_k)\|^2$$

$$\Rightarrow f(\underline{x}_{k+1}) - f(\underline{x}^*) \le \frac{1}{2s}\left[\|\underline{x}_k - \underline{x}^*\|^2 - \|\underline{x}_{k+1} - \underline{x}^*\|^2\right). \quad (5)$$

step ④: Summing (5) from 1 to k.

$$\sum_{i=1}^{k}\left(f(\underline{x}_i) - f(\underline{x}^*)\right) \le \frac{1}{2s}\left(\|\underline{x}_0 - \underline{x}^*\|^2 - \|\underline{x}_k - \underline{x}^*\|^2\right).$$

$$\le \frac{1}{2s}\|\underline{x}_0 - \underline{x}^*\|^2$$

Since $\{f(\underline{x}_k)\}$ is mono. non-incr. (GD descent prop), we have.

$$f(\underline{x}_k) - f(\underline{x}^*) \le \frac{1}{k}\sum_{i=1}^{k}[f(\underline{x}_i) - f(\underline{x}^*)] \le \frac{\|\underline{x}_0 - \underline{x}^*\|^2}{2sk} = O\left(\frac{1}{k}\right). \quad \blacksquare$$

"classical result: $O(\frac{1}{k})$ of GD".

(HW) prove: $\|\underline{x}_k - \underline{x}^*\| = o(\frac{1}{k})$

Convergence Rate Analysis (Convex): Fixed Step Size

Proof Sketch.

- (**Descent Lemma**): ∇f is L-Lipschitz $\Rightarrow$

$$f(\mathbf{y}) \le f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) + \frac{L}{2}\|\mathbf{y} - \mathbf{x}\|_2^2, \quad \forall \mathbf{x}, \mathbf{y}$$

- Plugging in $\mathbf{x}_{k+1} = \mathbf{x}_k - s\nabla f(\mathbf{x}_k)$ to obtain:

$$f(\mathbf{x}_{k+1}) \le f(\mathbf{x}_k) - \left(1 - \frac{Ls}{2}\right) s\|\nabla f(\mathbf{x}_k)\|_2^2$$

- Using the convexity of f and taking $0 < s \le 1/L$, and , we have

$$\begin{aligned} f(\mathbf{x}_{k+1}) &\le f(\mathbf{x}^*) + \nabla f(\mathbf{x}_k)^\top (\mathbf{x}_k - \mathbf{x}^*) - \frac{s}{2}\|\nabla f(\mathbf{x}_k)\|_2^2 \\ &= f(\mathbf{x}^*) + \frac{1}{2s}\left(\|\mathbf{x}_k - \mathbf{x}^*\|_2^2 - \|\mathbf{x}_{k+1} - \mathbf{x}^*\|_2^2\right) \end{aligned}$$

Convergence Rate Analysis (Convex): Fixed Step Size

- Summing over iterations & after telescoping:

$$\sum_{i=1}^{k} \big(f(\mathbf{x}_i) - f(\mathbf{x}^*)\big) \leq \frac{1}{2s}\big(\|\mathbf{x}_0 - \mathbf{x}^*\|_2^2 - \|\mathbf{x}_k - \mathbf{x}^*\|_2^2\big)$$
$$\leq \frac{1}{2s}\|\mathbf{x}_0 - \mathbf{x}^*\|_2^2$$

- Since $f(\mathbf{x}_k)$ is non-increasing, we have

$$f(\mathbf{x}_k) - f(\mathbf{x}^*) \leq \frac{1}{k}\sum_{i=1}^{k}\big(f(\mathbf{x}_i) - f(\mathbf{x}^*)\big) \leq \frac{\|\mathbf{x}_0 - \mathbf{x}^*\|_2^2}{2sk}. \qquad \square$$

Convergence Rate Analysis (Nonconvex): Fixed Step Size

Assume that f is nonconvex & differentiable, and L-smooth

Theorem 2 (Stationarity Gap)

If f is nonconvex, differentiable, and L-smooth, then gradient descent with fixed step size $s \leq 1/L$ satisfies

$$\min_{t=0,\ldots,k-1} \|\nabla f(\mathbf{x}_t)\|_2^2 \leq \frac{2(f(\mathbf{x}_0) - f^*)}{sk}$$

i.e., gradient descent method has sublinear convergence rate $O(1/k)$.

Remark:

- To get $\|\nabla f(\mathbf{x}_k)\|_2 \leq \epsilon$ for some k, it takes $O(\epsilon^{-2})$ iterations.

Theorem 2 (Stationarity Gap)

If f is nonconvex, differentiable, and L-smooth, then gradient descent with fixed step size $s \le 1/L$ satisfies

$$\min_{t=0,\ldots,k-1} \|\nabla f(\mathbf{x}_t)\|_2^2 \le \frac{2(f(\mathbf{x}_0) - f^*)}{sk}$$

i.e., gradient descent method has sublinear convergence rate $O(1/k)$.

Proof. We know that:

$$\left.\begin{aligned} & f(x_{k+1}) \le f(x_k) - s\left(1 - \frac{Ls}{2}\right)\|\nabla f(x_k)\|^2. \\ & \text{Also, } 0 < s \le \frac{1}{L} \Rightarrow -s\left(1-\frac{Ls}{2}\right) \le -\frac{s}{2} \end{aligned}\right\} \Rightarrow$$

$$f(x_{k+1}) - f(x_k) \le -\frac{s}{2}\|\nabla f(x_k)\|^2 \qquad (1)$$

Summing (1) from 0 to $k-1$:

$$\underbrace{f(x_k) - f(x_0)}_{\ge f(x^*) - f(x_0).} \le -\frac{s}{2}\sum_{t=0}^{k-1}\|\nabla f(x_t)\|^2 \le -\frac{sk}{2}\min_{t=0,\ldots,k-1}\|\nabla f(x_k)\|^2$$

Let $f^* = \inf_{x\in\mathbb{R}^n} f(x) > -\infty$. Then:

$$\min_{t=0,\ldots,k-1}\|\nabla f(x_t)\|^2 \le \frac{2(f(x_0)-f^*)}{sk} = O\left(\frac{1}{k}\right). \quad \square$$

Strategy 2: Exact Line Search

Choose the step size s to do the "best" we can along the direction of $-\nabla f(\mathbf{x})$:

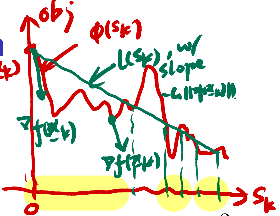
$$s = \arg\min_{t \geq 0} \underbrace{f(\mathbf{x}_k - t\nabla f(\mathbf{x}_k))}_{\text{dir fn.}}$$

Limitations:

$$-\nabla f(x_{k+1})^T \nabla f(x_k) = -\nabla f(x_k - t\nabla f(x_k))^T \nabla f(x_k) = 0$$

- Usually it's too expensive to do this in each iteration.

Strategy 3: Inexact Line Search



Seek s_k that satisfies Wolfe conditions:

- "Sufficient decrease" in f:

$$f(\mathbf{x}_{k+1}) = \underbrace{f(\mathbf{x}_k - s_k \nabla f(\mathbf{x}_k))}_{=\phi(s_k)} \le \underbrace{f(\mathbf{x}_k) - c_1 s_k \|\nabla f(\mathbf{x}_k)\|^2}_{L(s_k)}, \quad (0 < c_1 \ll 1)$$

- "Not zigzagging too badly":

$$-\nabla f(\mathbf{x}_{k+1})^\top \nabla f(\mathbf{x}_k) \le -c_2 \|\nabla f(\mathbf{x}_k)\|^2, \quad (c_1 < c_2 < 1)$$

$\ge$

$\nabla f(x_{k+1})$

$\nabla f(x_k)$

Main features:

- Can show that accumulation points $\bar{\mathbf{x}}$ of $\{\mathbf{x}_k\}$ are stationary: $\nabla f(\bar{\mathbf{x}}) = \mathbf{0}$ (thus minimizer if f is convex)
- Can do 1-dim line search for s_k, taking minima of quadratic or cubic interpolations of f and ∇f at the last two values tried. Use brackets for reliability. Often finds suitable s_k within 3 attempts (see [Nocedal & Wright, 2006, Ch. 3])

Strategy 3: Inexact Line Search – Backtracking

One way to adaptively choose step size is to use backtracking line search

1. First fix parameters $0 < \beta < 1$ and $0 < \alpha \le \frac{1}{2}$
2. At each iteration, start with $s = 1$, and while

$$f(\mathbf{x} - s\nabla f(\mathbf{x})) > f(\mathbf{x}) - \alpha s \|\nabla f(\mathbf{x})\|_2^2$$

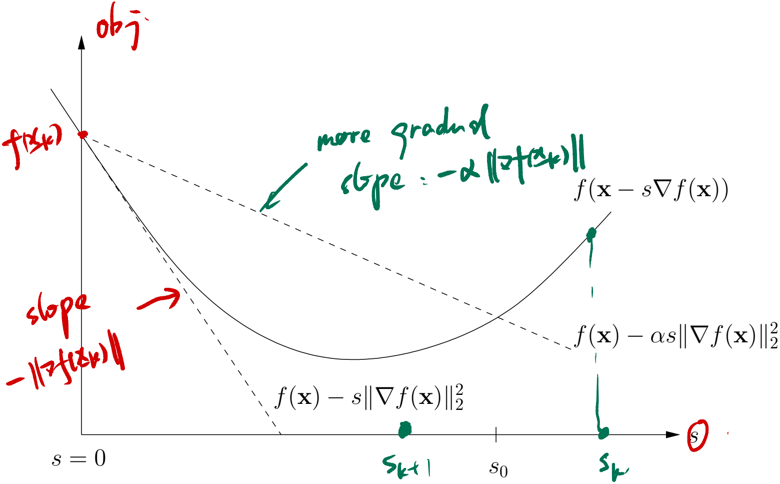
shink $\underbrace{s = \beta s}_{\text{shrinking}}$. Else, perform gradient descent update:

$$\mathbf{x}^+ = \mathbf{x} - s\nabla f(\mathbf{x})$$

Remarks:

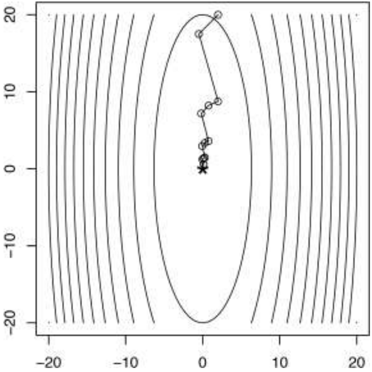
- Simple and tends to work well in practice (further simplification: just take $\alpha = \beta = 1/2$). But doesn't work for f nonsmooth
- Also referred to as Armijo's rule. Step size shrinking very aggressively
- Not checking the second Wolfe condition: the s_k thus identified is "within striking distance" of an s that's not too large

Backtracking Interpretation



Backtracking Example

Backtracking picks up roughly the right step size (12 outer iterations, 40 iterations in total):



$O(\frac{1}{k})$.

Next Class

Stochastic Gradient Descent