

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 $(\mathbf{x}_k, \mathbf{x}_{k-1}, \dots)$.

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)

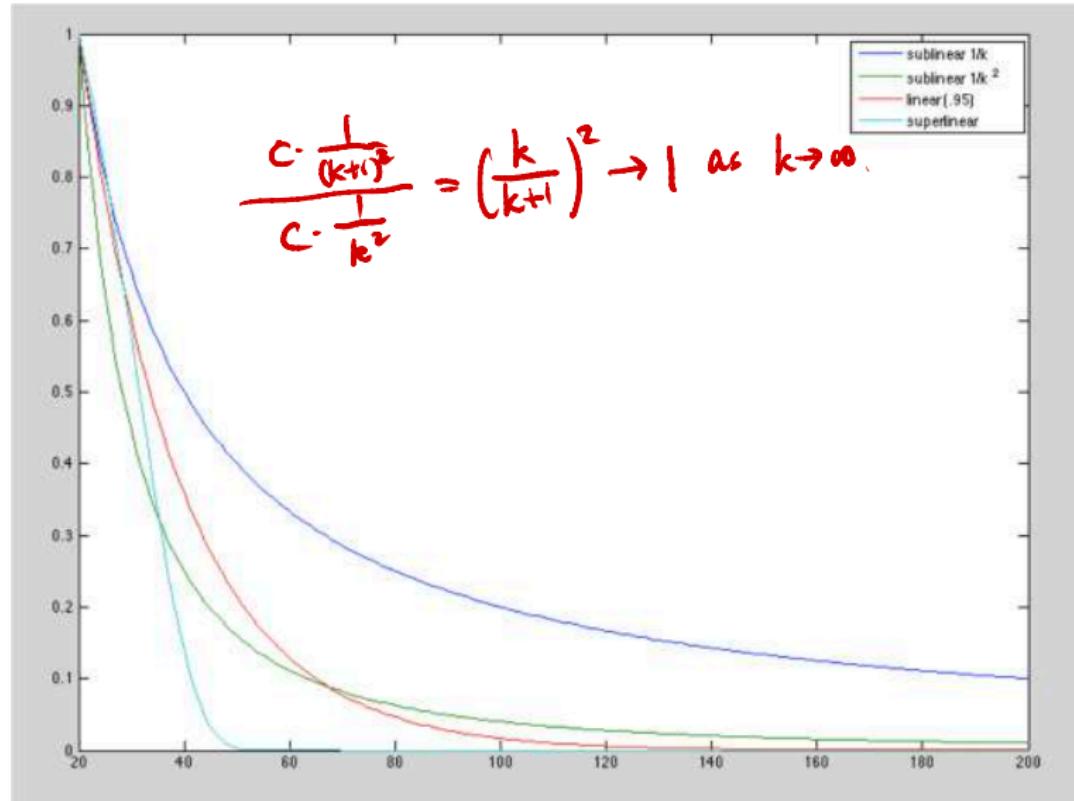
A sequence $\{r_k\} \rightarrow r^*$ and $r_k \neq r^*$ for all k . The rate (or order) of convergence p is a nonnegative number satisfying

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

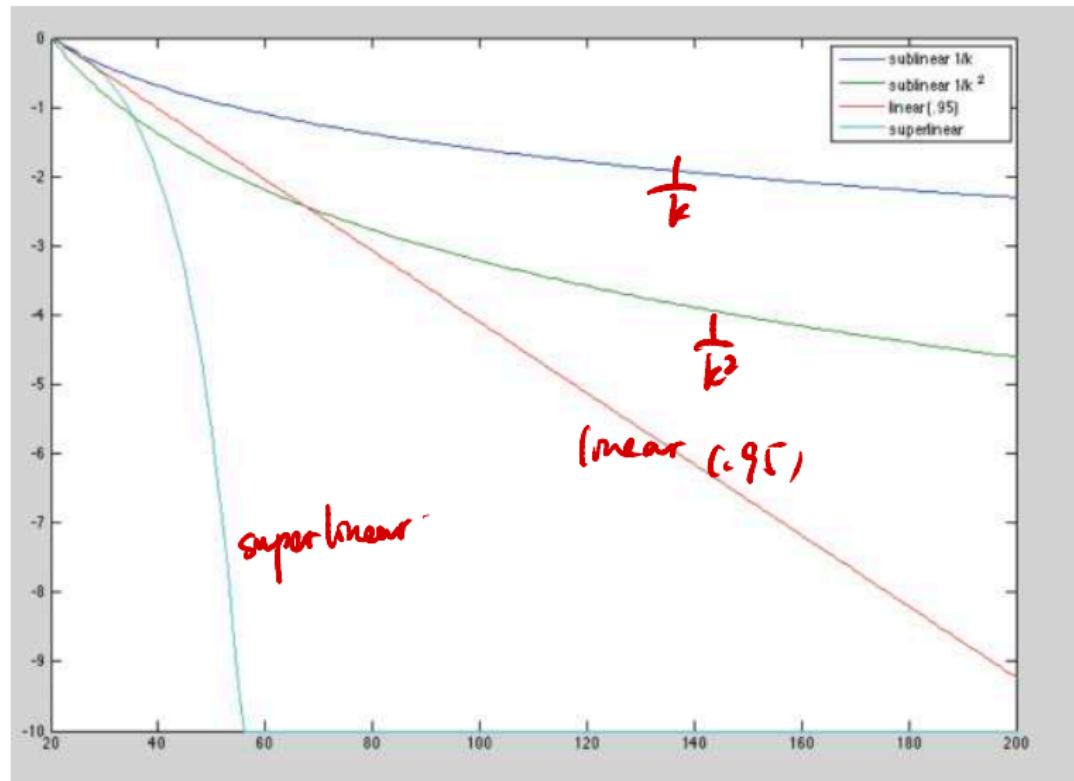
$$\frac{\|r_{k+1} - r^*\|}{\|r_k - r^*\|} \rightarrow 1 \text{ as } k \rightarrow \infty.$$

- Sublinear: $p = 1$ and $\beta = 1$ (e.g., $O(1/k)$ rate, kind of slow but still OK)
 $\frac{c(r_{k+1})}{c(r_k)} = \frac{k}{k+1} \rightarrow 1$. Desired $\varepsilon > 0$, $\frac{C}{k} \leq \varepsilon \rightarrow k \geq \frac{C}{\varepsilon} = O(\frac{C}{\varepsilon})$ contraction.
- Linear or geometric: $p = 1$ and $0 < \beta < 1$ (i.e., $\|r_{k+1} - r^*\| \leq \beta \|r_k - r^*\|$ for some $\beta \in (0, 1)$, or $\|r_k - r^*\| = O(\beta^k)$, which is quite fast)
Desired ε : $C\beta^k \leq \varepsilon \Rightarrow k \geq C \log(\varepsilon^{-1})$. Need $O(\log(\varepsilon^{-1}))$ iter.
- Superlinear: $p > 1$ and $\beta < \infty$, or $p = 1$ and $\beta = 0$ (i.e., $\frac{\|r_{k+1} - r^*\|}{\|r_k - r^*\|} \rightarrow 0$, that's very fast!) Not only a contraction, but also the rate of convergence is accelerating.
- Quadratic: $p = 2$ and $\beta < \infty$ ($\|r_{k+1} - r^*\| \leq \beta \|r_k - r^*\|^2$, # 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



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:

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

Stop if some stopping criterion is satisfied.

$$\text{e.g., } \|\nabla f(\mathbf{x}_k)\| \leq \varepsilon$$

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

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

Gradient Descent: Geometric Interpretation

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

$$f(y) \approx f(x) + \nabla f(x)^\top (y - x) + \frac{1}{2} (y - x)^\top \nabla^2 f(x) (y - x) + o(\|y - x\|)$$

so-approx *F0-approx* $\frac{1}{s} I$

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

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

F0-approx "proximity penalty"

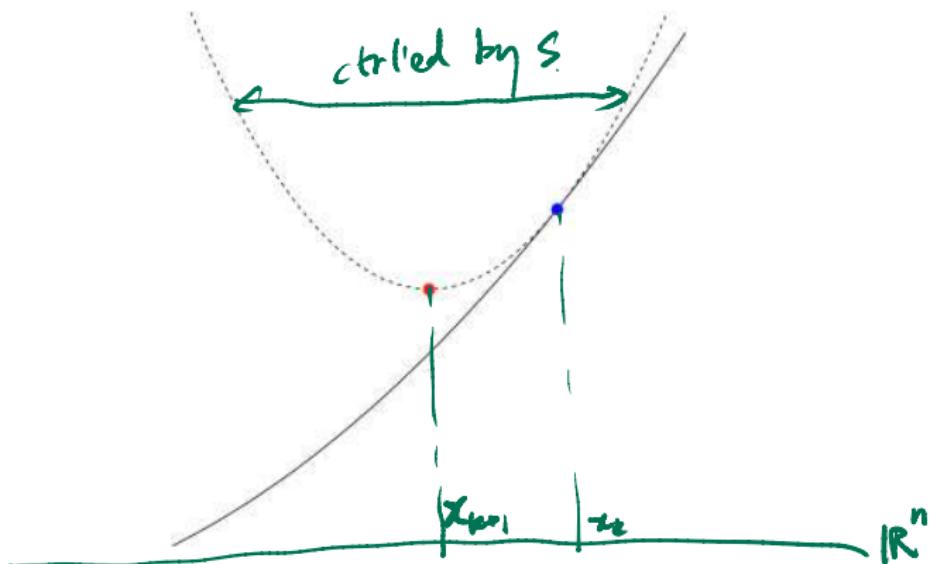
Can be viewed as a linear approximation to f , with proximity term to x weighted by $\frac{1}{2s}$. Choose next point $y = x^+$ to minimize this approximation:

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

Quadratic fn of y : set grad $\rightarrow 0$, solve for y

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

Gradient Descent: Geometric Interpretation



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

Questions:

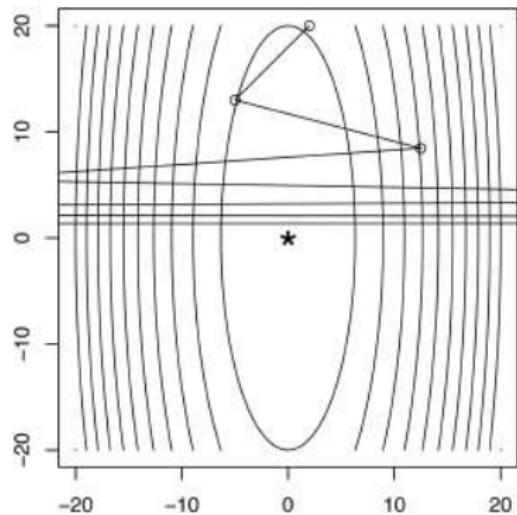
- How to choose step sizes $\{s_k\}$?
- What is the according convergence rate? Or does it depend on $\{s_k\}$?

Strategy 1: Fixed Step Size

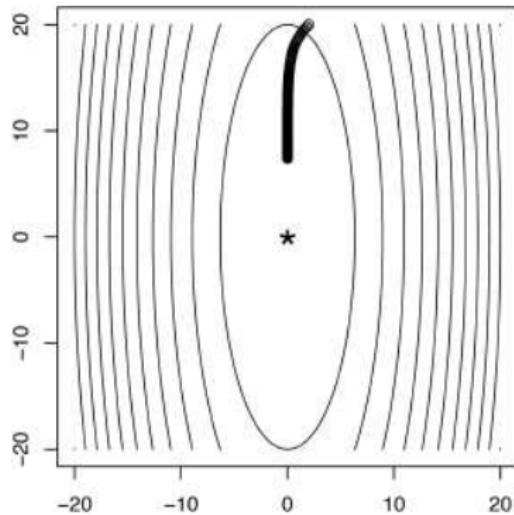
Simply set $s_k = s$ for all $k = 1, 2, 3, \dots$

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

Example: Consider $f(\mathbf{x}) = (10x_1^2 + x_2^2)/2$: $\Rightarrow (\mathbf{x}_1^*, \mathbf{x}_2^*) = (0, 0)$.



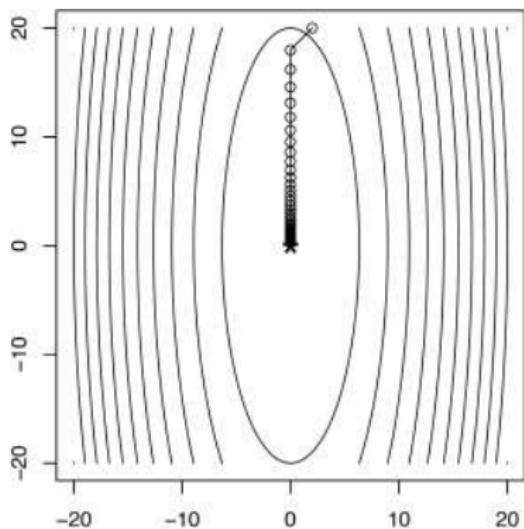
8 iterations



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 $\text{dom}(f) = \mathbb{R}^n$ 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)

$h: \mathbb{R}^n \rightarrow \mathbb{R}$ is Lip. cont. $\exists L > 0$ st. $\|h(\mathbf{y}) - h(\mathbf{x})\| \leq L\|\mathbf{y} - \mathbf{x}\|$.

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 $s \leq 1/L$ 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 $O(1/k)$.

Proof. Step 0 claim: If f is Lipschitz.. then descent lemma

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

To show (1), we consider

dir. der.

$$f(y) = f(x) + \int_0^1 f'(x + \tau(y-x)) d\tau.$$

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

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

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

By rearranging & taking abs. val. on both sides:

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

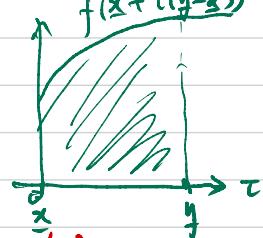
$$\leq \int_0^1 |[\nabla f(x + \tau(y-x)) - \nabla f(x)]^T (y-x)| d\tau. \quad (\text{Triangle Ineq.})$$

$$\quad \quad \quad \|a+b\| \leq \|a\| + \|b\|$$

$$\leq \int_0^1 \|\nabla f(x + \tau(y-x)) - \nabla f(x)\| \cdot \|y-x\| d\tau. \quad (\text{Cauchy-Schwarz})$$

$$\quad \quad \quad |a^T b| \leq \|a\| \cdot \|b\|$$

L -Lipschitz: $\leq L \tau \|y-x\|$



$$\leq \int_0^1 L \tau \|y - z\|^2 d\tau = L \|y - z\|^2 \underbrace{\int_0^1 \tau d\tau}_{= \frac{1}{2}} = \frac{L}{2} \|y - z\|^2.$$

(1) is proved.

Step ②: WTS "Descent property of GD".

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

$$f(z_{k+1}) \leq f(z_k) + \nabla f(z_k)^T \underbrace{(z_{k+1} - z_k)}_{-s_k \nabla f(z_k)} + \frac{L}{2} \underbrace{\|z_{k+1} - z_k\|^2}_{-s_k \nabla f(z_k)}$$

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

General:

$$= f(z_k) - s(1 - \frac{Ls}{2}) \|\nabla f(z_k)\|^2 \leftarrow \text{starting pt.}$$

Convergence Rate Analysis (Convex): Fixed Step Size

Proof Sketch.

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

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

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

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

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

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

Convergence Rate Analysis (Convex): Fixed Step Size

- Summing over iterations & after telescoping:

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

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

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

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

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

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.

Strategy 2: Exact Line Search

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

$$s = \arg \min_{t \geq 0} f(\mathbf{x} - t \nabla f(\mathbf{x}))$$

Limitations:

- 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 :

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

- “Not zigzagging too badly”:

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

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 $0 < \alpha \leq \frac{1}{2}$
- ② At each iteration, start with $s = 1$, and while

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

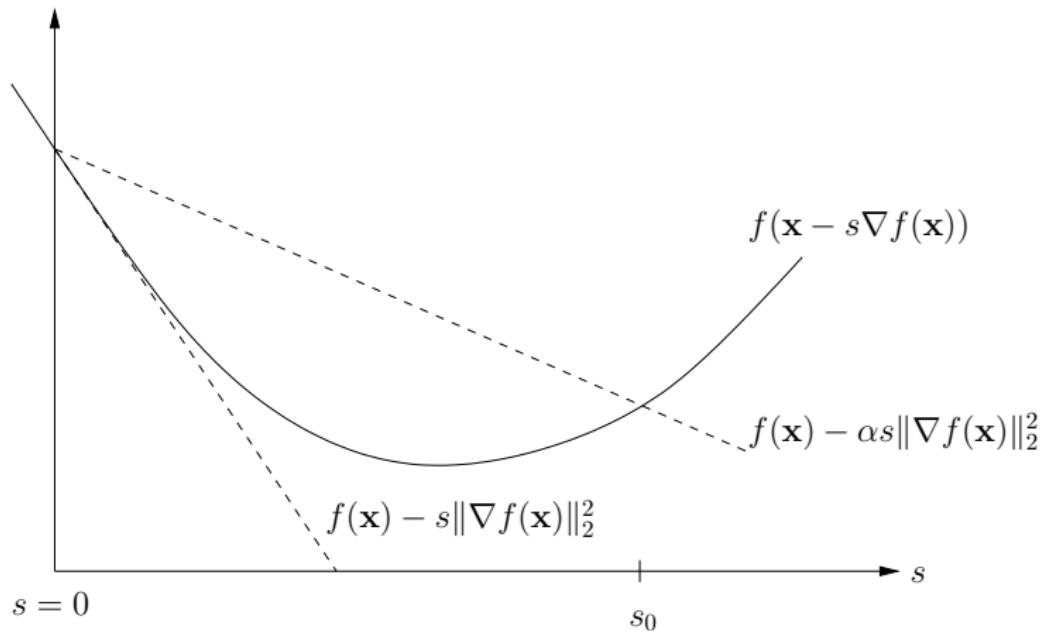
shrink $s = \beta s$. Else, perform gradient descent update:

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

Remarks:

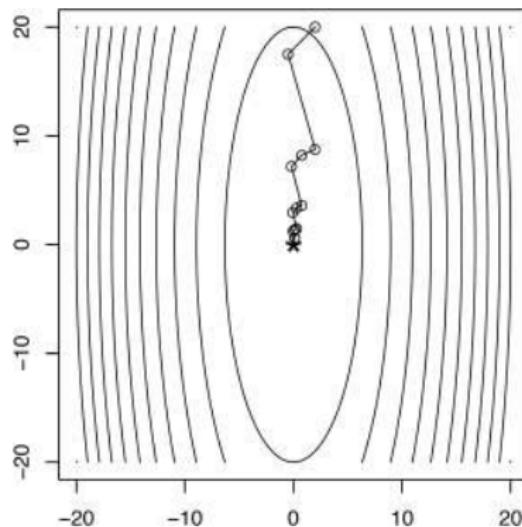
- Simple and tends to work well in practice (further simplification: just take $\alpha = \beta = 1/2$). But doesn't work for f nonsmooth
- Also referred to as **Armijo's rule**. Step size shrinking very aggressively
- Not checking the second Wolfe condition: the s_k thus identified is “within striking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



Next Class

Stochastic Gradient Descent