ECE 8101: Nonconvex Optimization for Machine Learning

Lecture Note 2-3: Gradient Descent

Jia (Kevin) Liu

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

Spring 2022

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?

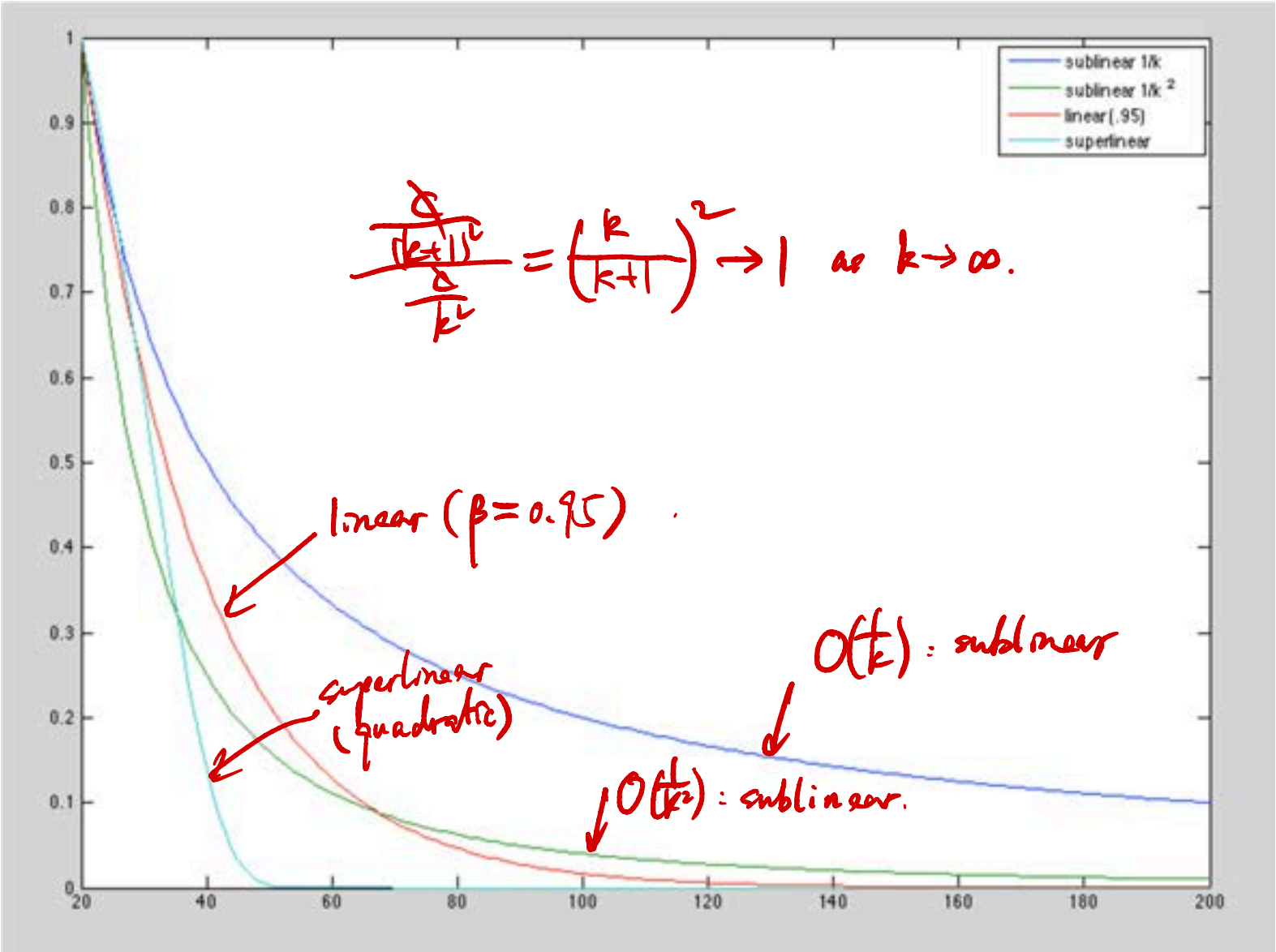
Definition 1 (Convergence rate)

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

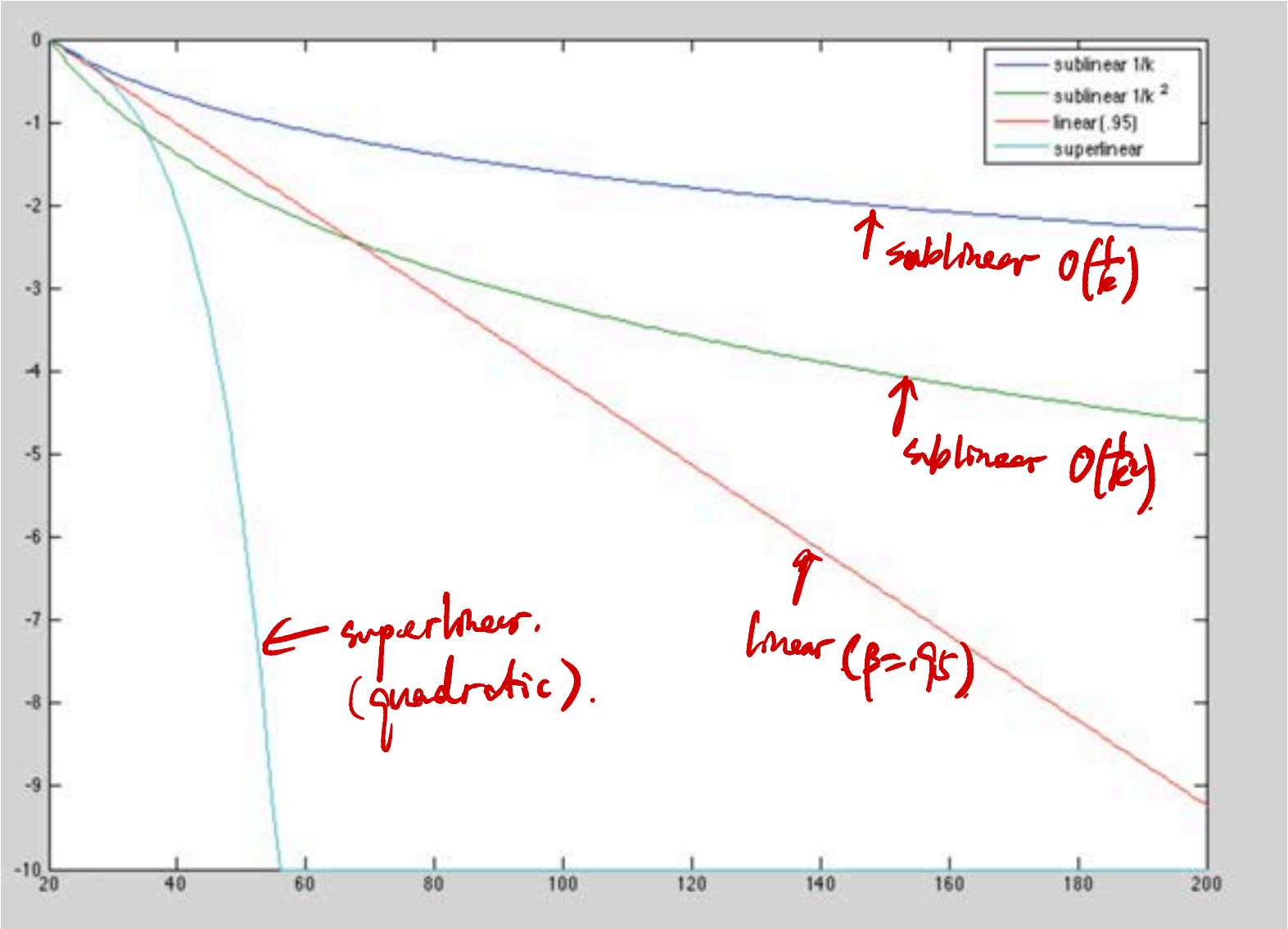
$$\limsup_{k\to\infty} \frac{\|r_{k+1} - r^*\|}{\|r_k - r^*\|^p} = \beta < \infty.$$

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

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:

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

Stop if some stopping criterion is satisfied.

(e.g., $\|\nabla f(x_k)\| \le \varepsilon$, $\|x_{k+1} - x_k\| \le \varepsilon$, some # of iters, ...).

Gradient Descent: Geometric Interpretation

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

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

SO-approx

FO-approx.

$+ o(\|\mathbf{y} - \mathbf{x}\|^2)$.

$\frac{1}{s}\mathbf{I}$

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

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

FO-approx.

"proximity term": penalize moving too far away from $\mathbf{x}$.

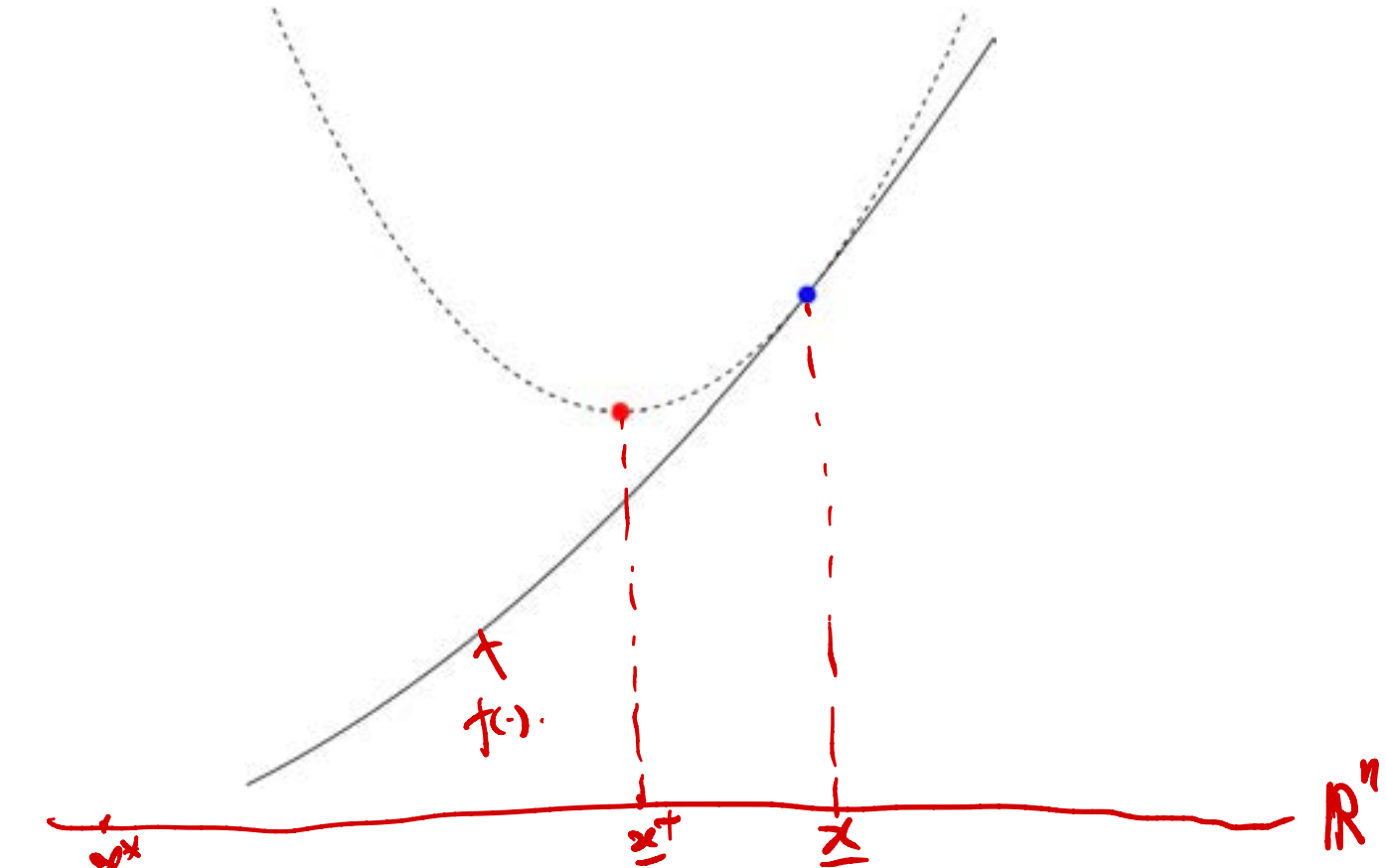
Can be viewed as a linear approximation to f, with proximity term to $\mathbf{x}$ weighted by $\frac{1}{2s}$. Choose next point $\mathbf{y} = \mathbf{x}^+$ to minimize this approximation:

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

Quad fn of y: unconstr, set grad to 0, then solve for y.

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

Gradient Descent: Geometric Interpretation



Questions:

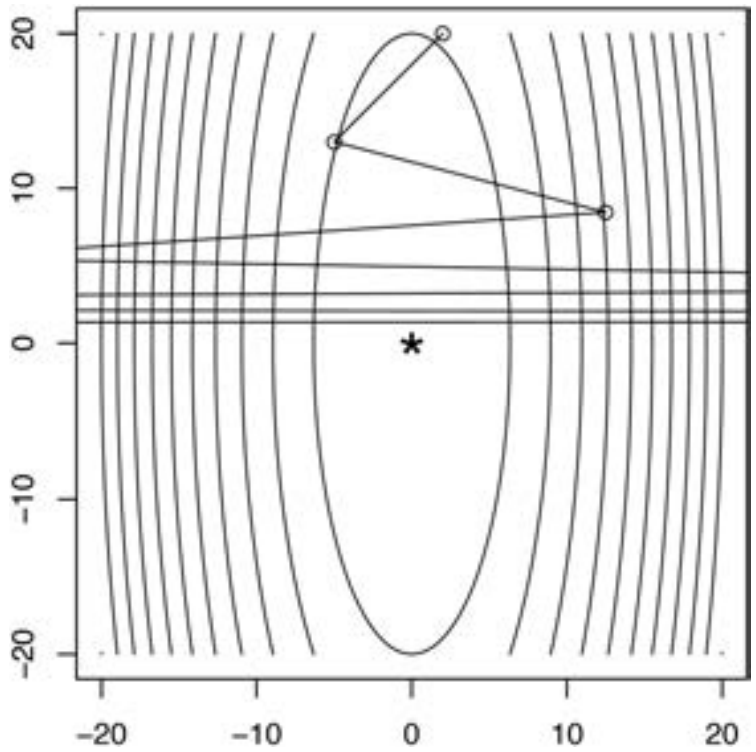
- How to choose step sizes $\{s_k\}$?
- What is the according convergence rate? Or does it depend on $\{s_k\}$?

Strategy 1: Fixed Step Size

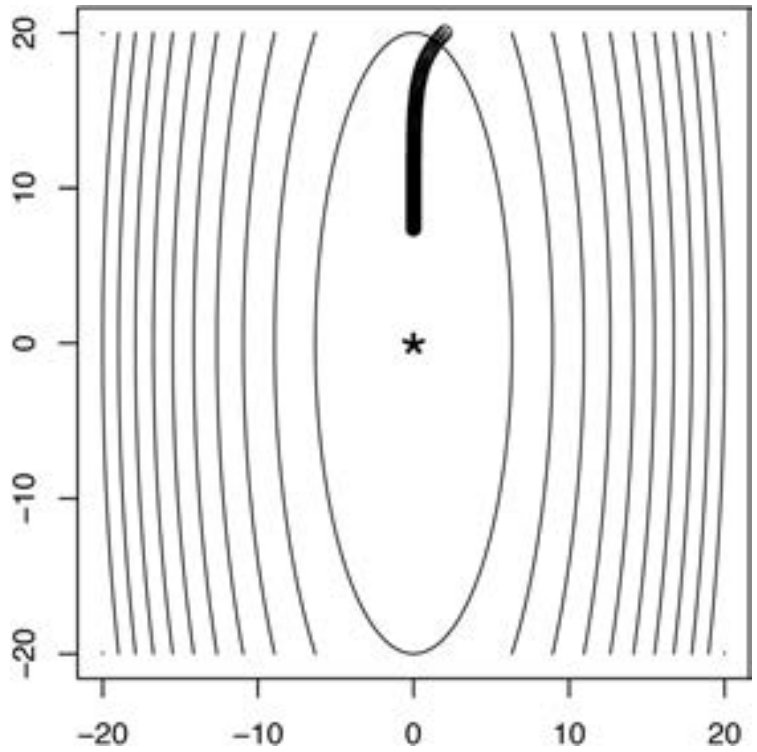
Simply set $s_k = s$ for all $k = 1, 2, 3, \ldots$.

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

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



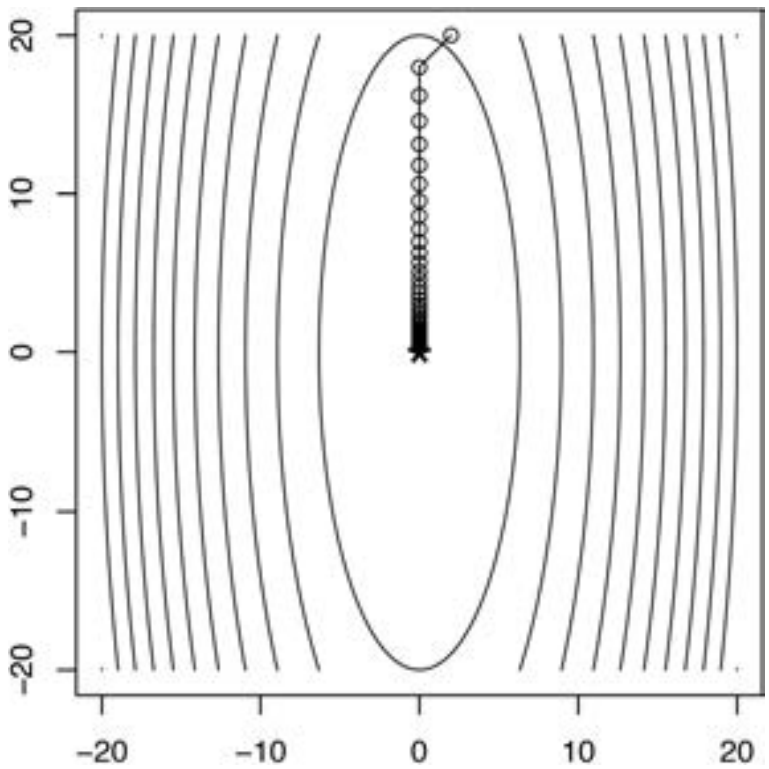
8 iterations



100 iterations

Strategy 1: Fixed Step Size

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



Dynamic syst. View Pt:

$\dot{x} = -\nabla f(x(t))$

Lyapunov fn:

$V(x) = f(x) - f(x^*)$.

$\frac{dV(x(t))}{dt} = \nabla f(x)^T \frac{dx(t)}{dt}$

$= -\nabla f(x(t))^T \nabla f(x(t))$

$= -\|\nabla f(x(t))\|_2^2 \leq 0$.

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

Need info of the "Lipschitz const." of $\nabla f(x)$ (smoothness).

Convergence Rate Analysis (Convex): Fixed Step Size

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

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

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

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

Theorem 1 (Optimality Gap)

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

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

init pt. dist.

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

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

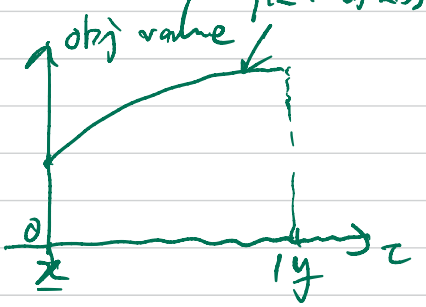
(init pt. dist.)

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

Proof: step 0: claim: If ∇f is Lipschitz, then:

$$f(y) \le f(x) + \nabla f(x)^T (y-x) + \frac{L}{2}\|y-x\|_2^2, \quad \forall x, y \in \mathbb{R}^n.$$

(descent lemma) (1).



To show (1), we start from: (dir. der.)

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

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

(add & subtract $\nabla f(x)^T(y-x)$)

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

By some rearranging and take abs. value on both sides:

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

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

(Triangle ineq: $\|a+b\| \le \|a\| + \|b\|$)

$$\le \int_0^1 \underbrace{\|\nabla f(x+\tau(y-x)) - \nabla f(x)\|}_{L\text{-Lipschitz. } \le L\tau\|y-x\|} \cdot \|y-x\| \cdot d\tau.$$

(Cauchy–Schwartz ineq: $|a^T b| \le \|a\| \cdot \|b\|$.)

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

(1) is proved.

Step ②: WTS: "Descent property of GD":

$x_{k+1} = x_k - s\,\nabla f(x_k)$. Plugging this in (1).

$$f(x_{k+1}) \le f(x_k) + \nabla f(x_k)^T \underbrace{(x_{k+1} - x_k)}_{= -s\nabla f(x_k)} + \frac{L}{2}\|x_{k+1} - x_k\|_2^2.$$

$$= f(x_k) - s\|\nabla f(x_k)\|_2^2 + \frac{Ls^2}{2}\|\nabla f(x_k)\|_2^2$$

$$= f(x_k) - s\left(1 - \frac{Ls}{2}\right)\|\nabla f(x_k)\|_2^2. \qquad (2)$$

Step ③: From convexity of $f(x)$, we have:

$$f(x^*) \ge f(x_k) + \nabla f(x_k)^T (x^* - x_k).$$

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

Plugging (3) into (2) yields:

$$f(x_{k+1}) \le f(x^*) + \nabla f(x_k)^T (x_k - x^*) - s\left(1 - \frac{Ls}{2}\right)\|\nabla f(x_k)\|_2^2 \qquad (4).$$

Recall step-size $s \in \left(0, \frac{1}{L}\right]$. Then,

$$0 < s \le \frac{1}{L} \Rightarrow 0 < Ls \le 1 \Rightarrow -\frac{1}{2} \le -\frac{Ls}{2} < 0 \Rightarrow \frac{1}{2} \le 1 - \frac{Ls}{2} < 1.$$

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

Use the above in (4) $\Rightarrow$

$$f(x_{k+1}) - f(x^*) \le \nabla f(x_k)^T (x_k - x^*) - \frac{s}{2}\|\nabla f(x_k)\|_2^2 \qquad (5).$$

step ④ : Consider RHS of (5):

$$\nabla f(x_k)^T(x_k - x^*) - \frac{s}{2}\|\nabla f(x_k)\|_2^2$$

$$= -\frac{1}{2s}\left[s^2\|\nabla f(x_k)\|_2^2 - 2s\nabla f(x_k)^T(x_k - x^*) + \|x_k - x^*\|_2^2 - \|x_k - x^*\|_2^2\right]$$

$$= -\frac{1}{2s}\left[\|(x_k - x^*) - s\nabla f(x_k)\|_2^2 - \|x_k - x^*\|_2^2\right]$$

$$= -\frac{1}{2s}\left[\|x_{k+1} - x^*\|_2^2 - \|x_k - x^*\|_2^2\right]$$

Therefore, (5) $\Rightarrow$

$$f(x_{k+1}) - f(x^*) \le \frac{1}{2s}\left(\|x_k - x^*\|_2^2 - \|x_{k+1} - x^*\|_2^2\right). \qquad (6).$$

step ⑤: Summing (6) from 0 to k-1 (telescoping).

$$\sum_{i=0}^{k-1}(f(x_i) - f(x^*)) \le \frac{1}{2s}\left(\|x_0 - x^*\|_2^2 - \|x_k - x^*\|_2^2\right).$$

$$\le \frac{1}{2s}\|x_0 - x^*\|_2^2$$

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

$$f(x_k) - f(x^*) \le \frac{1}{k}\sum_{i=1}^{k}[f(x_i) - f(x^*)] \le \frac{\|x_0 - x^*\|_2^2}{2sk} = O\left(\frac{1}{k}\right). \qquad \square$$

Classic $O\left(\frac{1}{k}\right)$ of GD.

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 $0 < s \leq 1/L$, and , we have

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

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

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

□

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

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

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 $s \leq 1/L$ satisfies

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

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

Proof: We know that,

$$\left.\begin{aligned} f(x_{k+1}) &\leq f(x_k) - s\left(1 - \frac{Ls}{2}\right)\|\nabla f(x_k)\|_2^2 \\ \text{Also, } 0 < s \leq \tfrac{1}{L} &\Rightarrow -s\left(1 - \frac{Ls}{2}\right) \leq -\frac{s}{2} \end{aligned}\right\} \Rightarrow$$

$f(x_{k+1}) \leq f(x_k) - \frac{s}{2}\|\nabla f(x_k)\|_2^2$. Summing from 0 to $k-1$:

$$f(x_k) - f(x_0) \leq -\frac{s}{2}\sum_{t=0}^{k-1}\|\nabla f(x_t)\|_2^2 \leq -\frac{sk}{2}\min_{t=0,\ldots,k-1}\|\nabla f(x_t)\|_2^2$$

$\geq f(x^*) - f(x_0)$

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

$$\min_{t=0,\ldots,k-1}\|\nabla f(x_t)\|_2^2 \leq \frac{2(f(x_0) - f^*)}{sk} = O(\tfrac{1}{k}).$$

Strategy 2: Exact Line Search

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

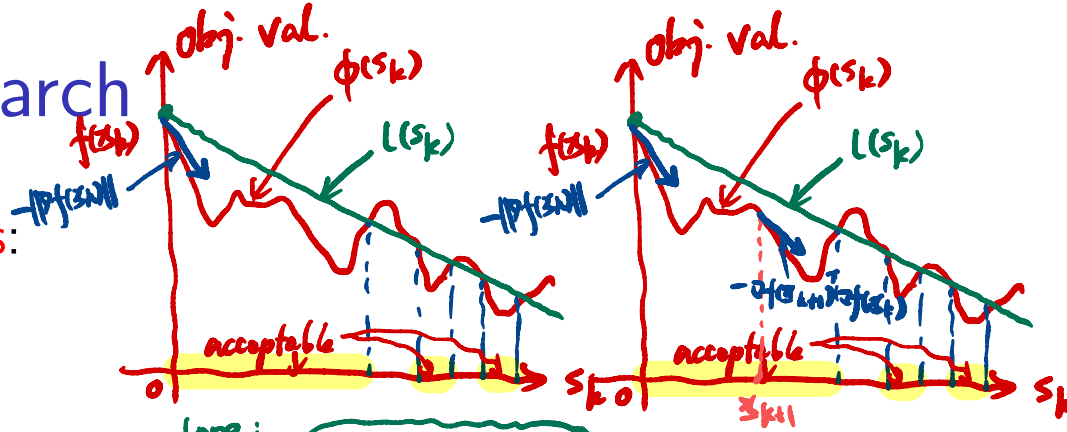
$$s = \arg\min_{t \geq 0} f(\mathbf{x} - t\nabla f(\mathbf{x}))$$

Limitations:

Take dir. der. w.r.t. t, set it to 0, solve for t. $-\nabla f(\mathbf{x}_{k+1})^T \nabla f(\mathbf{x}_k) = 0$

- 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:

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

- "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

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

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

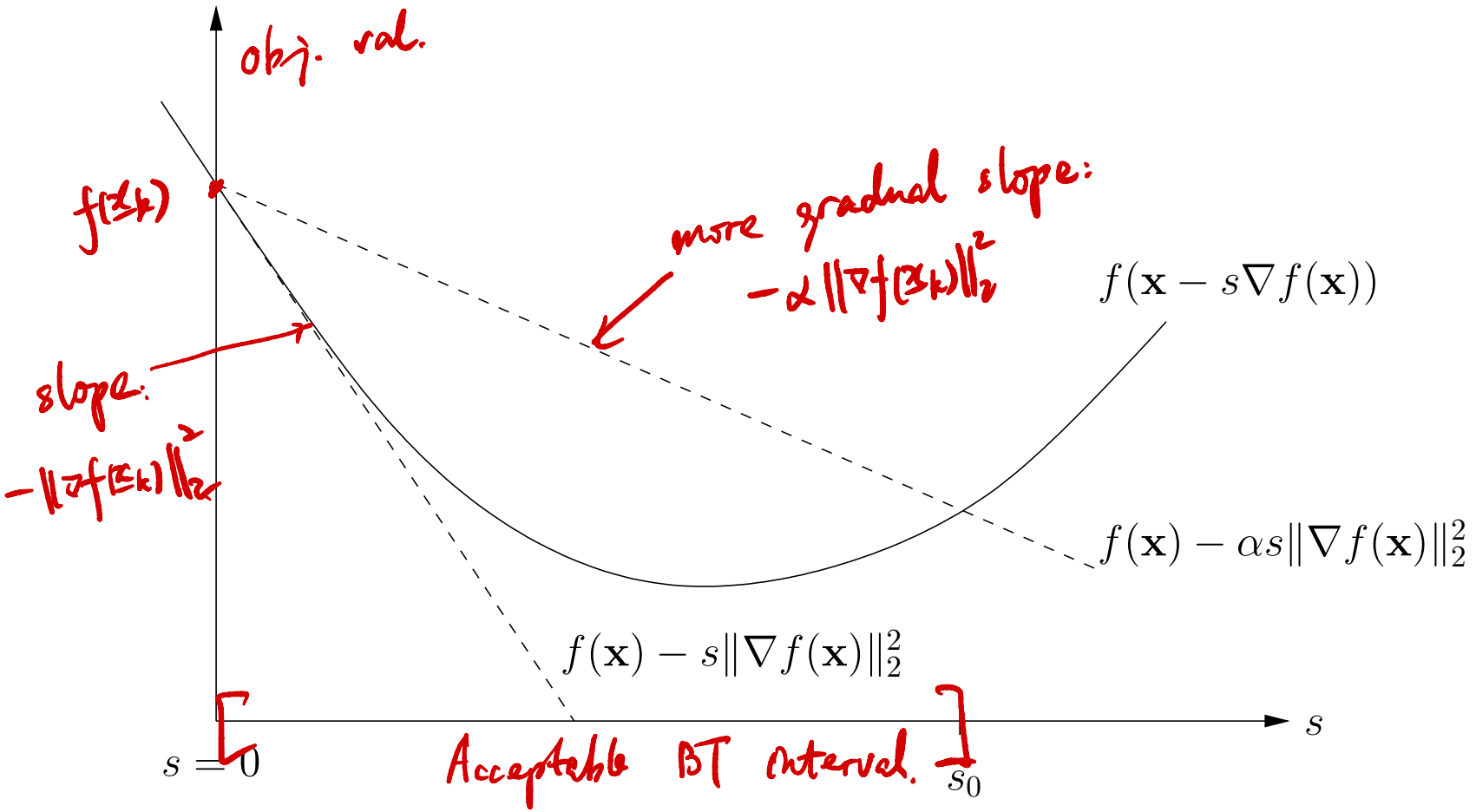
shrink by a β-factor.

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

Remarks:

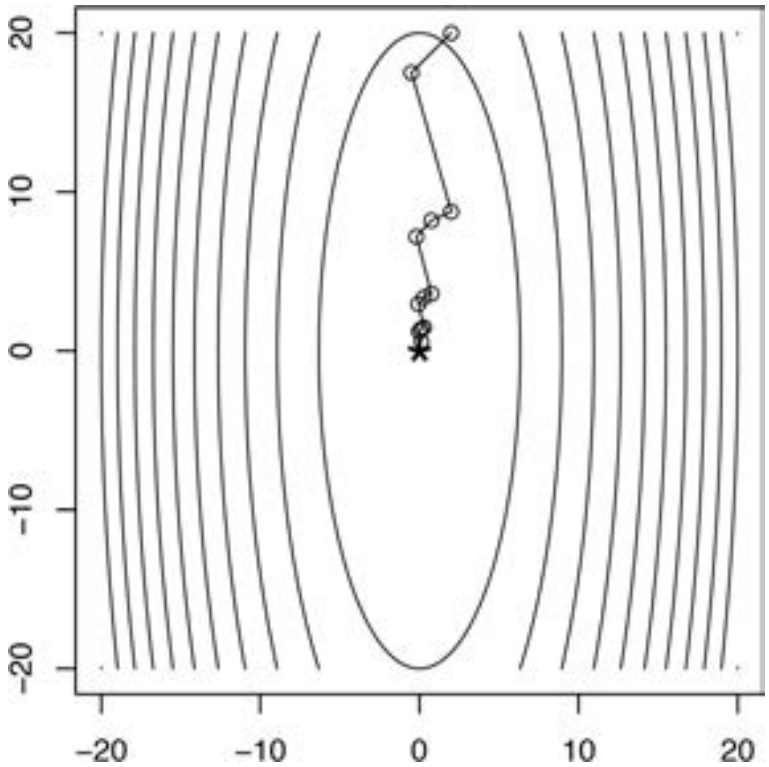
- Simple and tends to work well in practice (further simplification: just take $\alpha = \beta = 1/2$). But doesn't work for f nonsmooth
- Also referred to as Armijo's rule. Step size shrinking very aggressively
- Not checking the second Wolfe condition: the s_k thus identified is "within striking distance"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