Gradient Descent

Lecturer: Pradeep Ravikumar
Co-instructor: Aarti Singh

Convex Optimization 10-725/36-725

Based on slides from Vandenberghe, Tibshirani

Gradient Descent

Consider unconstrained, smooth convex optimization

$$\min_x \; f(x)$$

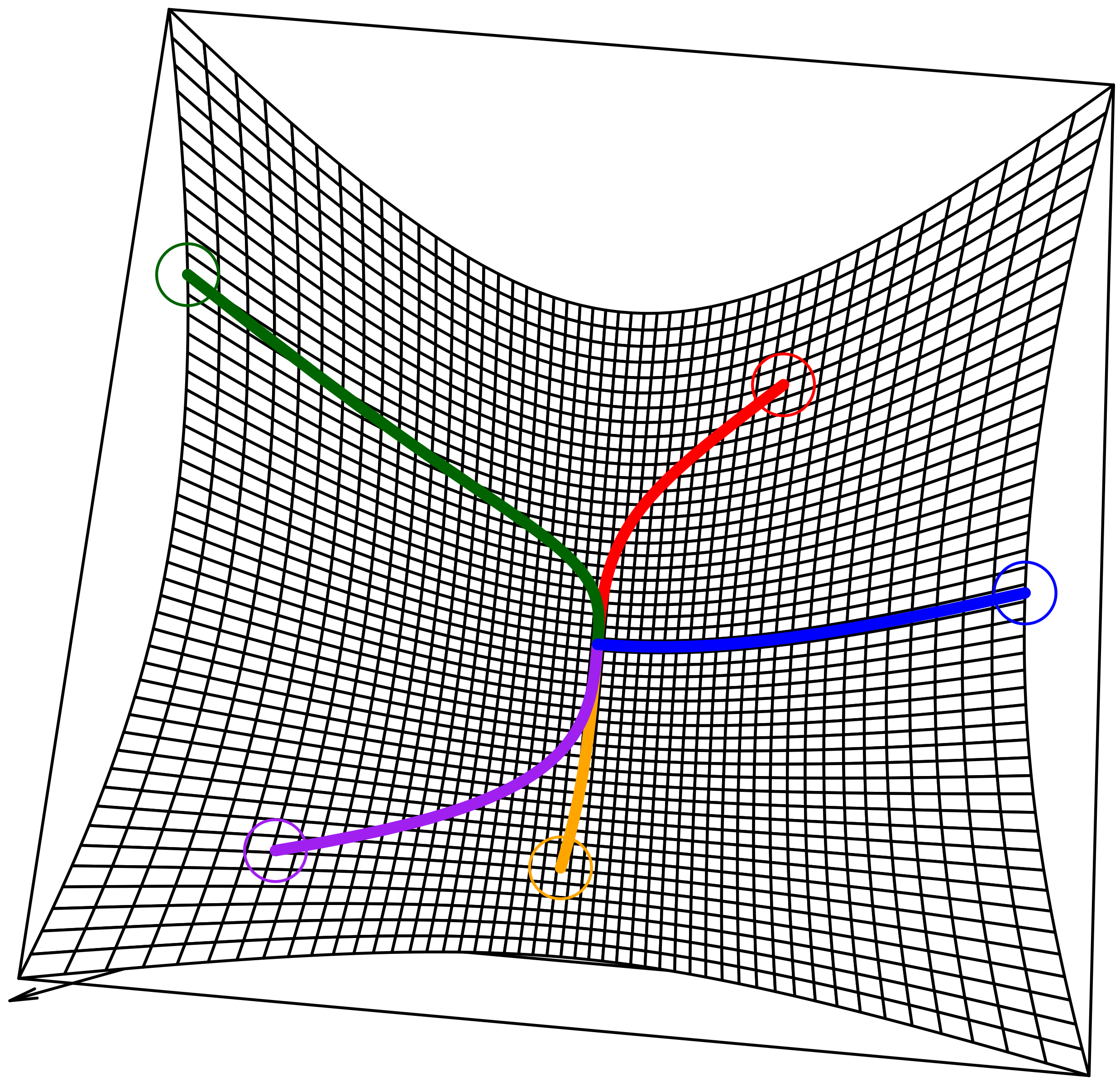
i.e., f is convex and differentiable with $\mathrm{dom}(f) = \mathbb{R}^n$. Denote the optimal criterion value by $f^\star = \min_x \; f(x)$, and a solution by $x^\star$

Gradient descent: choose initial point $x^{(0)} \in \mathbb{R}^n$, repeat:

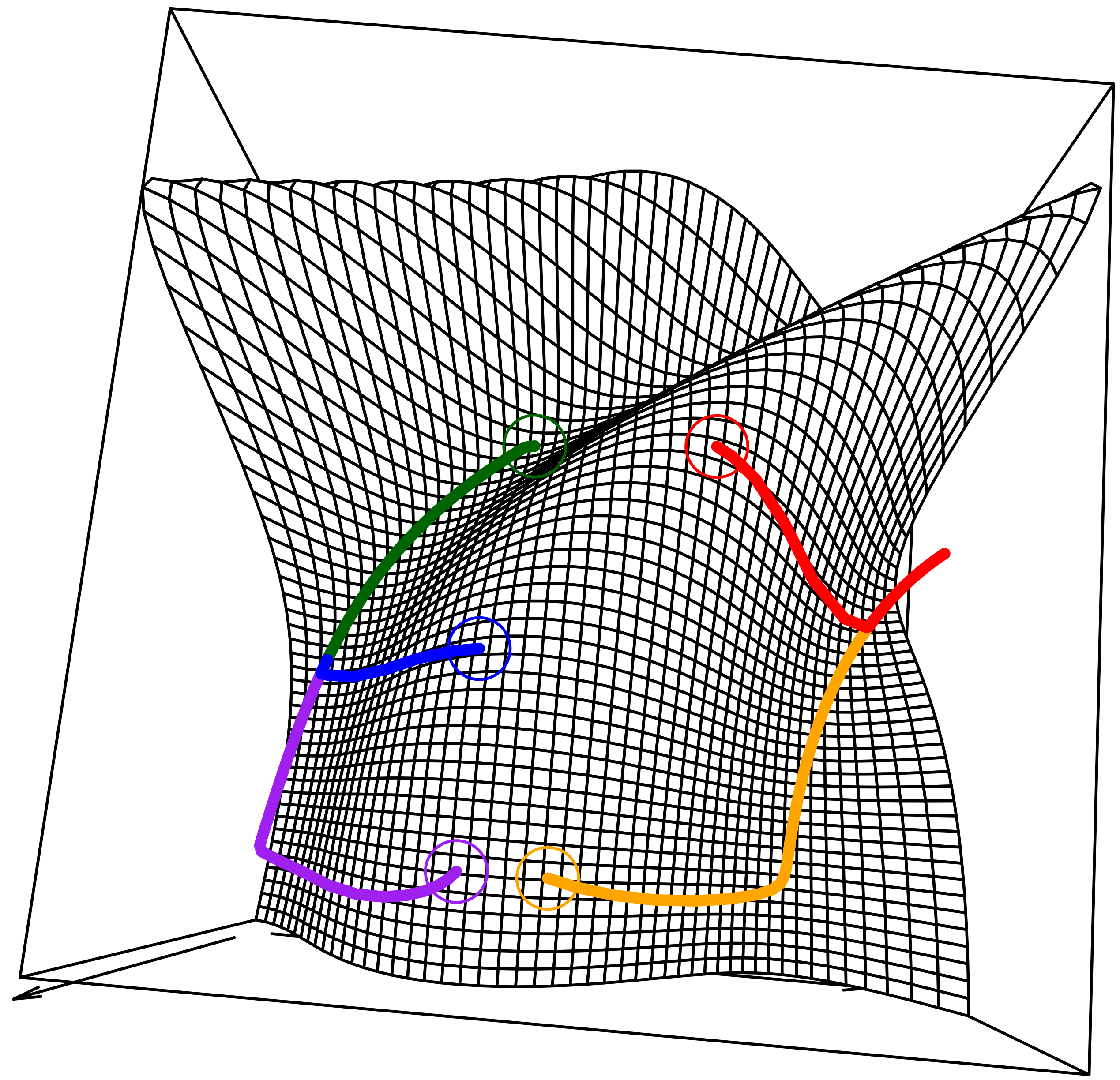
$$x^{(k)} = x^{(k-1)} - t_k \cdot \nabla f(x^{(k-1)}), \quad k = 1, 2, 3, \ldots$$

Stop at some point

Example I



Example II

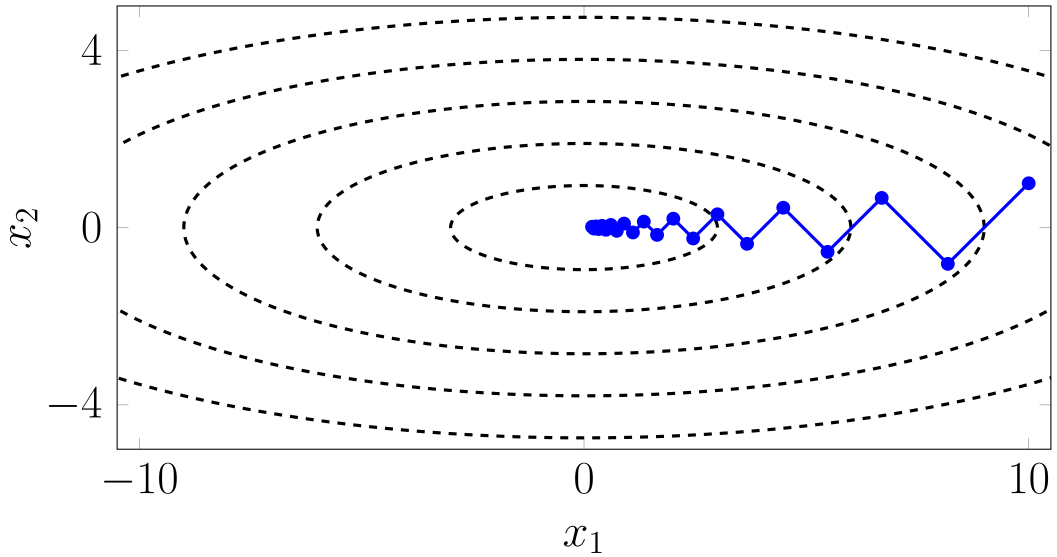


Quadratic Example

$$f(x) = \frac{1}{2}(x_1^2 + \gamma x_2^2) \qquad \text{(with } \gamma > 1\text{)}$$

with exact line search and starting point $x^{(0)} = (\gamma, 1)$

$$\frac{\|x^{(k)} - x^\star\|_2}{\|x^{(0)} - x^\star\|_2} = \left(\frac{\gamma - 1}{\gamma + 1}\right)^k$$

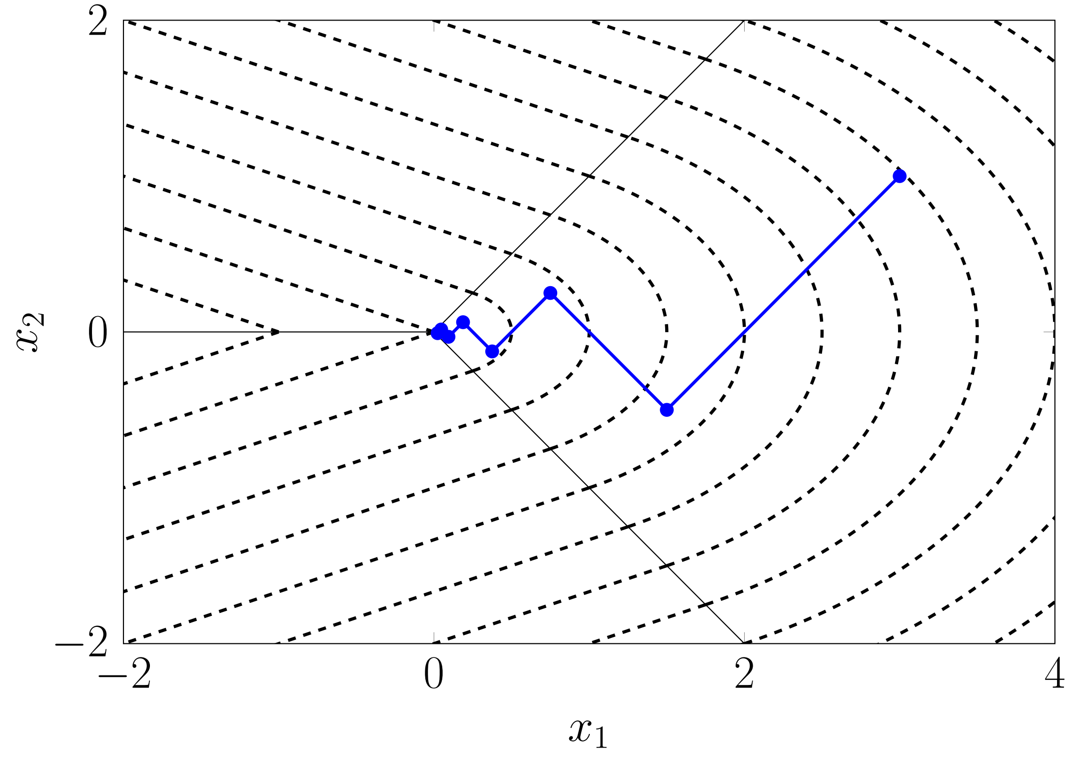


gradient method is often slow; convergence very dependent on scaling

Non-differentiable Example

$$f(x) = \sqrt{x_1^2 + \gamma x_2^2} \quad \text{for } |x_2| \le x_1, \qquad f(x) = \frac{x_1 + \gamma |x_2|}{\sqrt{1+\gamma}} \quad \text{for } |x_2| > x_1$$

with exact line search, starting point $x^{(0)} = (\gamma, 1)$, converges to non-optimal point



gradient method does not handle nondifferentiable problems

Descent-type algorithms with better guarantees

Methods with improved convergence

- quasi-Newton methods
- conjugate gradient method
- accelerated gradient method

Methods for nondifferentiable or constrained problems

- subgradient method
- proximal gradient method
- smoothing methods
- cutting-plane methods

Gradient Descent

- Now that we have seen how horrible gradient descent is, and how there are so many methods with better guarantees, let’s now go ahead and study gradient descent more closely

Gradient Descent

- Now that we have seen how horrible gradient descent is, and how there are so many methods with better guarantees, let's now go ahead and study gradient descent more closely
    - Why?

Gradient Descent

- Now that we have seen how horrible gradient descent is, and how there are so many methods with better guarantees, let’s now go ahead and study gradient descent more closely
  - Why?
- For unconstrained problems, gradient descent still empirically preferred (more robust, less tuning)

Gradient Descent

- Now that we have seen how horrible gradient descent is, and how there are so many methods with better guarantees, let’s now go ahead and study gradient descent more closely
  - Why?
- For unconstrained problems, gradient descent still empirically preferred (more robust, less tuning)
- For constrained, non-differentiable problems, algorithms are “variants” of gradient descent

Function Approximation Interpretation

At each iteration, consider the expansion

$$f(y) \approx f(x) + \nabla f(x)^T (y - x) + \frac{1}{2t} \|y - x\|_2^2$$

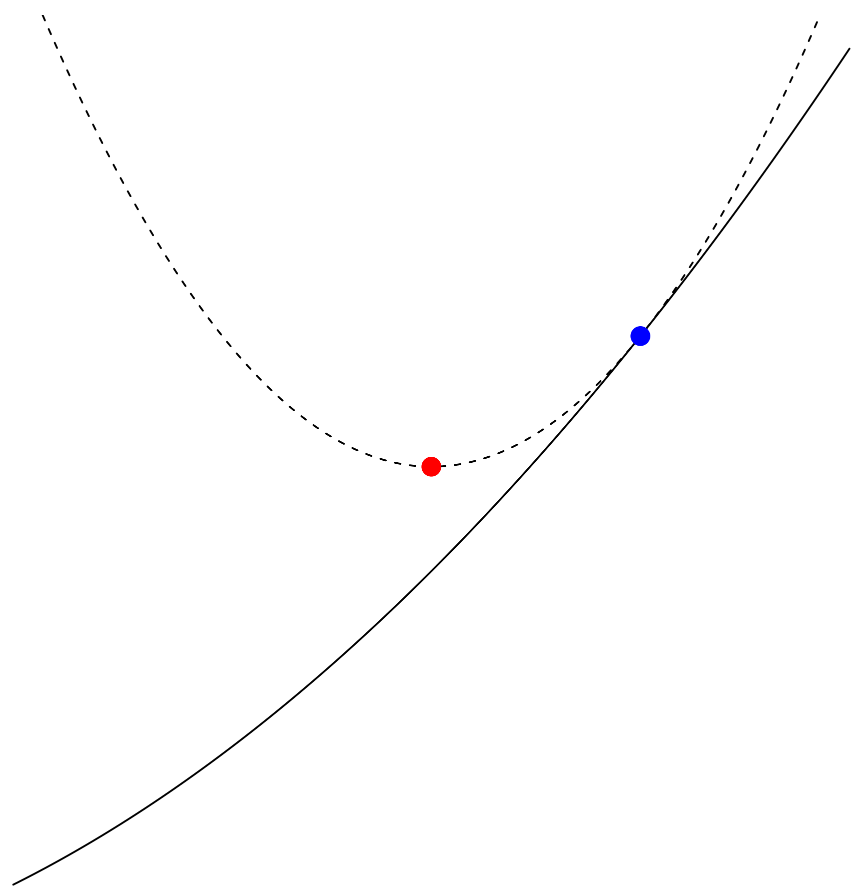
Quadratic approximation, replacing usual Hessian $\nabla^2 f(x)$ by $\frac{1}{t} I$

$f(x) + \nabla f(x)^T (y - x)$	linear approximation to f
$\frac{1}{2t} \|y - x\|_2^2$	proximity term to x, with weight $1/(2t)$

Choose next point $y = x^+$ to minimize quadratic approximation:

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

Function Approximation Interpretation



Blue point is x, red point is

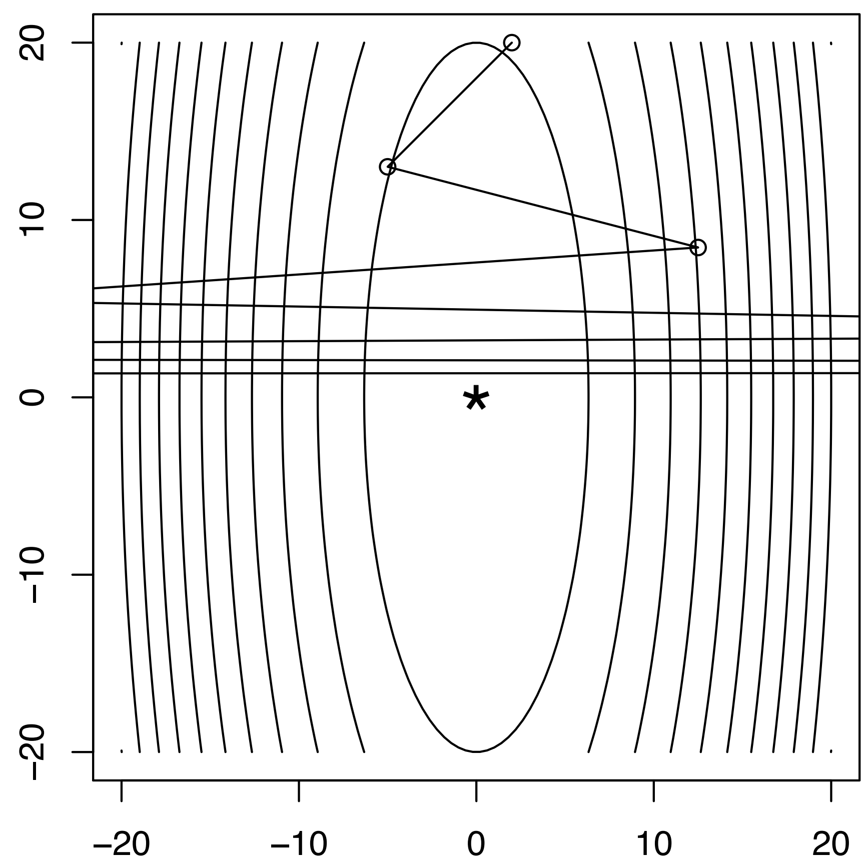
$$x^+ = \underset{y}{\operatorname{argmin}} \; f(x) + \nabla f(x)^T (y - x) + \frac{1}{2t}\|y - x\|_2^2$$

Gradient Descent

- How to choose step size
- Convergence Analysis

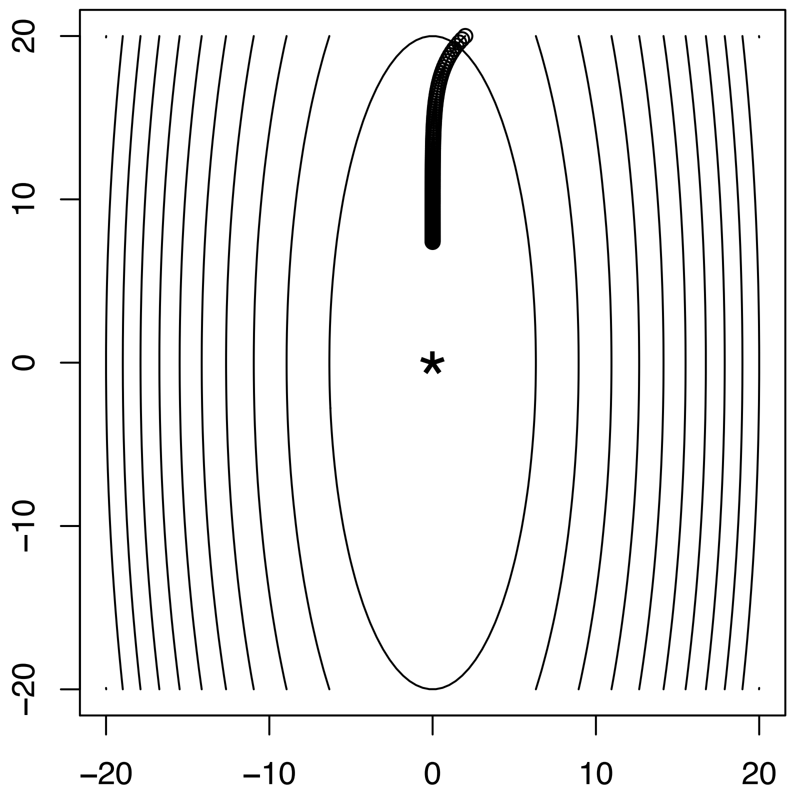
Fixed Step Size: Too Big

Simply take $t_k = t$ for all $k = 1, 2, 3, \ldots$, can diverge if t is too big.
Consider $f(x) = (10x_1^2 + x_2^2)/2$, gradient descent after 8 steps:



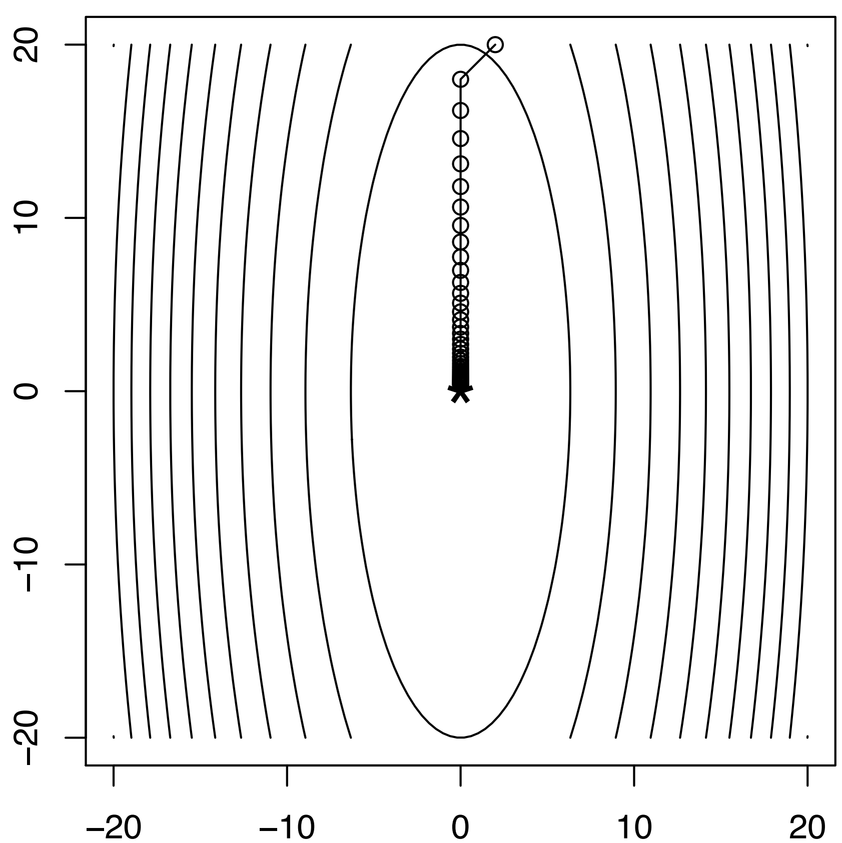
Fixed Step Size: Too Small

Can be slow if t is too small. Same example, gradient descent after 100 steps:



Fixed Step Size: Just Right

Converges nicely when t is "just right". Same example, gradient descent after 40 steps:



Convergence analysis later will give us a precise idea of "just right"

Step-Size: Backtracking Line Search

- First fix parameters $0 < \beta < 1$ and $0 < \alpha \leq 1/2$
- At each iteration, start with $t = t_{\text{init}}$, and while

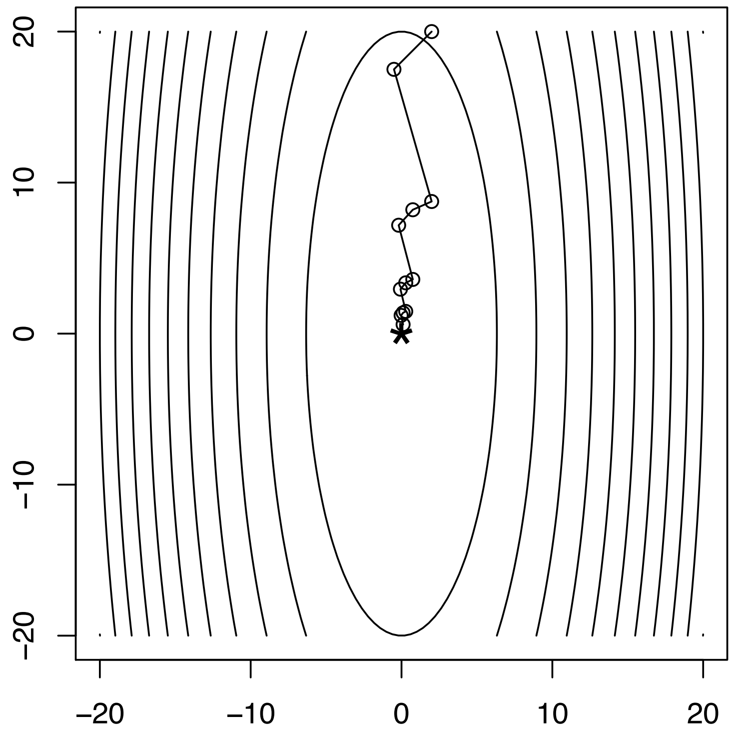
$$f(x - t\nabla f(x)) > f(x) - \alpha t \|\nabla f(x)\|_2^2$$

shrink $t = \beta t$. Else perform gradient descent update

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

Backtracking

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



Here $\alpha = \beta = 0.5$

Exact Line Search

Could also choose step to do the best we can along direction of negative gradient, called exact line search:

$$t = \underset{s \geq 0}{\operatorname{argmin}} \; f(x - s\nabla f(x))$$

Usually not possible to do this minimization exactly

Approximations to exact line search are often not as efficient as backtracking, and it's usually not worth it

Convergence Analysis: Convexity

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

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2 \quad \text{for any } x, y$$

I.e., ∇f is Lipschitz continuous with constant $L > 0$

Convergence Analysis: Convexity

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

$$\|\nabla f(x) - \nabla f(y)\|_2 \le L\|x - y\|_2 \quad \text{for any } x, y$$

I.e., ∇f is Lipschitz continuous with constant $L > 0$

Theorem: Gradient descent with fixed step size $t \le 1/L$ satisfies

$$f(x^{(k)}) - f^\star \le \frac{\|x^{(0)} - x^\star\|_2^2}{2tk}$$

Convergence Analysis: Convexity

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

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2 \quad \text{for any } x, y$$

I.e., ∇f is Lipschitz continuous with constant $L > 0$

Theorem: Gradient descent with fixed step size $t \leq 1/L$ satisfies

$$f(x^{(k)}) - f^\star \leq \frac{\|x^{(0)} - x^\star\|_2^2}{2tk}$$

We say gradient descent has convergence rate $O(1/k)$

I.e., to get $f(x^{(k)}) - f^\star \leq \epsilon$, we need $O(1/\epsilon)$ iterations

Proof

Key steps:

- ∇f Lipschitz with constant $L \Rightarrow$

$$f(y) \le f(x) + \nabla f(x)^T (y - x) + \frac{L}{2}\|y - x\|_2^2 \quad \text{all } x, y$$

Proof

Key steps:

- ∇f Lipschitz with constant $L \Rightarrow$

$$f(y) \le f(x) + \nabla f(x)^T (y - x) + \frac{L}{2}\|y - x\|_2^2 \quad \text{all } x, y$$

- Plugging in $y = x^+ = x - t\nabla f(x)$,

$$f(x^+) \le f(x) - \Big(1 - \frac{Lt}{2}\Big) t \|\nabla f(x)\|_2^2$$

Proof

Key steps:

- ∇f Lipschitz with constant $L \Rightarrow$

$$f(y) \le f(x) + \nabla f(x)^T (y-x) + \frac{L}{2}\|y-x\|_2^2 \quad \text{all } x, y$$

- Plugging in $y = x^+ = x - t\nabla f(x)$,

$$f(x^+) \le f(x) - \Big(1 - \frac{Lt}{2}\Big) t \|\nabla f(x)\|_2^2$$

- Taking $0 < t \le 1/L$, and using convexity of f,

$$\begin{aligned} f(x^+) &\le f^\star + \nabla f(x)^T (x - x^\star) - \frac{t}{2}\|\nabla f(x)\|_2^2 \\ &= f^\star + \frac{1}{2t}\big(\|x - x^\star\|_2^2 - \|x^+ - x^\star\|_2^2\big) \end{aligned}$$

Proof Contd.

$$f(x^{(i)}) - f^* \leq \frac{1}{2t}\left(\|x^{(i-1)} - x^*\|_2^2 - \|x^{(i)} - x^*\|_2^2\right)$$

Proof Contd.

- Summing over iterations:

$$\sum_{i=1}^{k}(f(x^{(i)}) - f^\star) \le \frac{1}{2t}\big(\|x^{(0)} - x^\star\|_2^2 - \|x^{(k)} - x^\star\|_2^2\big)$$
$$\le \frac{1}{2t}\|x^{(0)} - x^\star\|_2^2$$

- Since $f(x^{(k)})$ is nonincreasing,

$$f(x^{(k)}) - f^\star \le \frac{1}{k}\sum_{i=1}^{k}\big(f(x^{(i)}) - f^\star\big) \le \frac{\|x^{(0)} - x^\star\|_2^2}{2tk}$$

$\square$

Convergence Analysis: Backtracking

Same assumptions, f is convex and differentiable, $\mathrm{dom}(f) = \mathbb{R}^n$, and ∇f is Lipschitz continuous with constant $L > 0$

Same rate for a step size chosen by backtracking search

Theorem: Gradient descent with backtracking line search satisfies

$$f(x^{(k)}) - f^\star \leq \frac{\|x^{(0)} - x^\star\|_2^2}{2t_{\mathsf{min}}k}$$

where $t_{\mathsf{min}} = \min\{1, \beta/L\}$

If β is not too small, then we don't lose much compared to fixed step size (β/L vs $1/L$)

Convergence Analysis: Strong Convexity

Reminder: strong convexity of f means $f(x) - \frac{\mathrm{m}}{2}\|x\|_2^2$ is convex for some $m > 0$. If f is twice differentiable, then this is equivalent to

$$f(y) \geq f(x) + \nabla f(x)^T (y - x) + \frac{m}{2}\|y - x\|_2^2 \quad \text{all } x, y$$

Under Lipschitz assumption as before, and also strong convexity:

Theorem: Gradient descent with fixed step size $t \leq 2/(m + L)$ or with backtracking line search search satisfies

$$f(x^{(k)}) - f^\star \leq c^k \frac{L}{2}\|x^{(0)} - x^\star\|_2^2$$

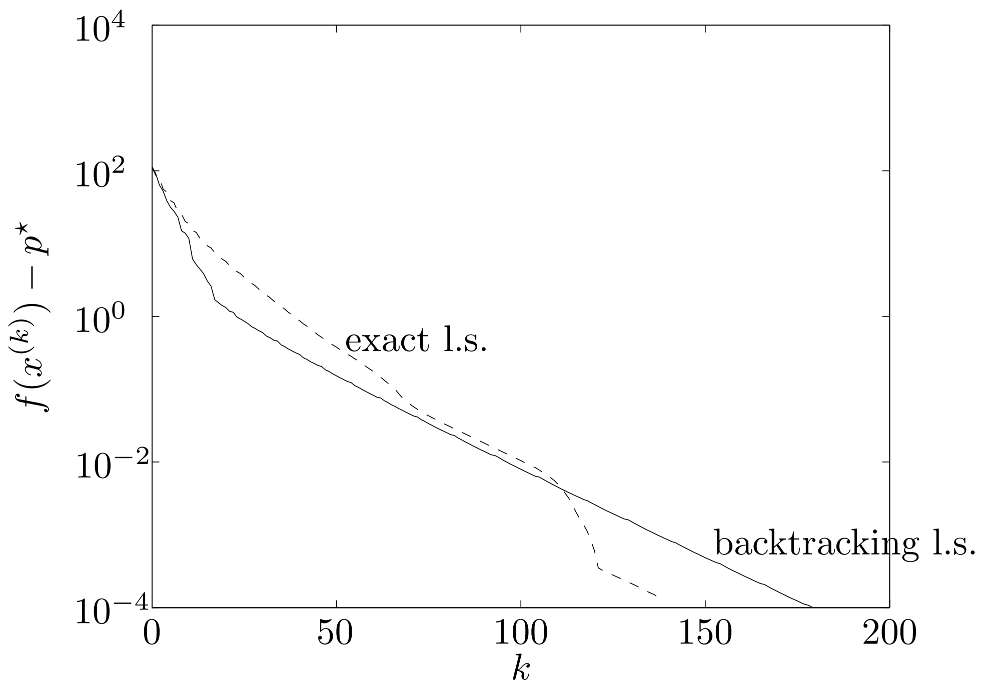
where $0 < c < 1$

Linear Convergence

I.e., rate with strong convexity is $O(c^k)$, exponentially fast!

I.e., to get $f(x^{(k)}) - f^\star \le \epsilon$, need $O(\log(1/\epsilon))$ iterations

Called linear convergence, because looks linear on a semi-log plot



(From B & V page 487)

Constant c depends adversely on condition number L/m (higher condition number $\Rightarrow$ slower rate)

A look at the conditions so far

A look at the conditions for a simple problem, $f(\beta) = \frac{1}{2}\|y - X\beta\|_2^2$

Lipschitz continuity of ∇f:

- This means $\nabla^2 f(x) \preceq LI$
- As $\nabla^2 f(\beta) = X^T X$, we have $L = \sigma^2_{\mathsf{max}}(X)$

Strong convexity of f:

- This means $\nabla^2 f(x) \succeq mI$
- As $\nabla^2 f(\beta) = X^T X$, we have $m = \sigma^2_{\mathsf{min}}(X)$
- If X is wide—i.e., X is $n \times p$ with $p > n$—then $\sigma_{\mathsf{min}}(X) = 0$, and f can't be strongly convex
- Even if $\sigma_{\mathsf{min}}(X) > 0$, can have a very large condition number $L/m = \sigma^2_{\mathsf{max}}(X)/\sigma^2_{\mathsf{min}}(X)$

A look at the conditions so far

A function f having Lipschitz gradient and being strongly convex satisfies:

$$mI \preceq \nabla^2 f(x) \preceq LI \quad \text{for all } x \in \mathbb{R}^n,$$

for constants $L > m > 0$

Think of f being sandwiched between two quadratics

A look at the conditions so far

A function f having Lipschitz gradient and being strongly convex satisfies:

$$mI \preceq \nabla^2 f(x) \preceq LI \quad \text{for all } x \in \mathbb{R}^n,$$

for constants $L > m > 0$

Think of f being sandwiched between two quadratics

May seem like a strong condition to hold globally (for all $x \in \mathbb{R}^n$). But a careful look at the proofs shows that we only need Lipschitz gradients/strong convexity over the sublevel set

$$S = \{x : f(x) \leq f(x^{(0)})\}$$

This is less restrictive (especially if S is compact)

Practicalities

Stopping rule: stop when $\|\nabla f(x)\|_2$ is small

- Recall $\nabla f(x^\star) = 0$ at solution $x^\star$
- If f is strongly convex with parameter m, then

$$\|\nabla f(x)\|_2 \leq \sqrt{2m\epsilon} \implies f(x) - f^\star \leq \epsilon \text{ (L/m)}$$

Pros and cons of gradient descent:

- Pro: simple idea, and each iteration is cheap (usually)
- Pro: fast for well-conditioned, strongly convex problems
- Con: can often be slow, because many interesting problems aren't strongly convex or well-conditioned
- Con: can't handle nondifferentiable functions

Can we do better?

Gradient descent has $O(1/\epsilon)$ convergence rate over problem class of convex, differentiable functions with Lipschitz gradients

First-order method: iterative method, updates $x^{(k)}$ in

$$x^{(0)} + \mathrm{span}\{\nabla f(x^{(0)}), \nabla f(x^{(1)}), \ldots \nabla f(x^{(k-1)})\}$$

Theorem (Nesterov): For any $k \leq (n-1)/2$ and any starting point $x^{(0)}$, there is a function f in the problem class such that any first-order method satisfies

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

Can attain rate $O(1/k^2)$, or $O(1/\sqrt{\epsilon})$? Answer: yes (we'll see)!

Proof: Convergence Analysis for Strong Convexity

Analysis for constant step size

if $x^+ = x - t\nabla f(x)$ and $0 < t \leq 2/(m+L)$:

$$
\begin{aligned}
\|x^+ - x^\star\|_2^2 &= \|x - t\nabla f(x) - x^\star\|_2^2 \\
&= \|x - x^\star\|_2^2 - 2t\nabla f(x)^T(x - x^\star) + t^2\|\nabla f(x)\|_2^2
\end{aligned}
$$

Proof: Convergence Analysis for Strong Convexity

Analysis for constant step size

if $x^+ = x - t\nabla f(x)$ and $0 < t \leq 2/(m+L)$:

$$\begin{aligned} \|x^+ - x^\star\|_2^2 &= \|x - t\nabla f(x) - x^\star\|_2^2 \\ &= \|x - x^\star\|_2^2 - 2t\nabla f(x)^T(x - x^\star) + t^2\|\nabla f(x)\|_2^2 \end{aligned}$$

$f(x)$ is m-strongly convex, and with L-Lipshitz gradients

$$\Rightarrow \ (\nabla f(x) - \nabla f(y))^T (x - y) \geq \frac{mL}{m+L}\|x - y\|_2^2 + \frac{1}{m+L}\|\nabla f(x) - \nabla f(y)\|_2^2$$

$$\Rightarrow \nabla f(x)^T(x - x^*) \geq \frac{mL}{m+L}\|x - x^*\|_2^2 + \frac{1}{m+L}\|\nabla f(x)\|_2^2$$

Proof: Convergence Analysis for Strong Convexity

Analysis for constant step size

if $x^+ = x - t\nabla f(x)$ and $0 < t \le 2/(m+L)$:

$$\begin{aligned}
\|x^+ - x^\star\|_2^2 &= \|x - t\nabla f(x) - x^\star\|_2^2 \\
&= \|x - x^\star\|_2^2 - 2t\nabla f(x)^T(x - x^\star) + t^2\|\nabla f(x)\|_2^2 \\
&\le (1 - t\frac{2mL}{m+L})\|x - x^\star\|_2^2 + t(t - \frac{2}{m+L})\|\nabla f(x)\|_2^2 \\
&\le (1 - t\frac{2mL}{m+L})\|x - x^\star\|_2^2
\end{aligned}$$

Proof Contd.

Distance to optimum

$$\|x^{(k)} - x^\star\|_2^2 \le c^k \|x^{(0)} - x^\star\|_2^2, \qquad c = 1 - t\frac{2mL}{m+L}$$

- implies (linear) convergence
- for $t = 2/(m+L)$, get $c = \left(\frac{\gamma - 1}{\gamma + 1}\right)^2$ with $\gamma = L/m$

Bound on function value

$$f(x^{(k)}) - f^\star \le \frac{L}{2}\|x^{(k)} - x^\star\|_2^2 \le \frac{c^k L}{2}\|x^{(0)} - x^\star\|_2^2$$