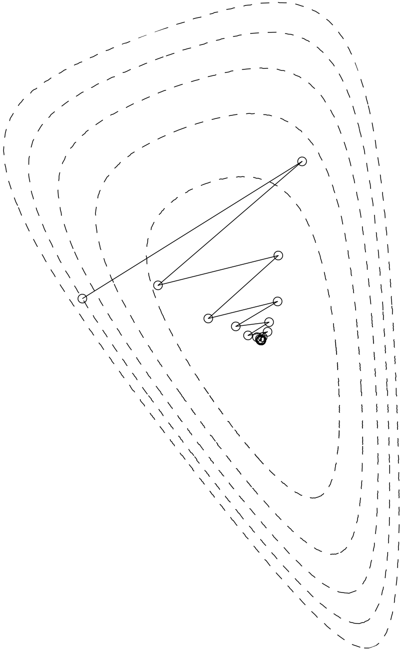


Optimization Algorithms

Gradient Descent & Backtracking Line Search

plain gradient descent, stepsize adaptation, backtracking line search, Wolfe conditions, exponential convergence

Marc Toussaint
Technical University of Berlin

Winter 2024/25

Gradient descent

- Problem: $\min_{x \in \mathbb{R}^n} f(x)$ for smooth objective function: $f : \mathbb{R}^n \to \mathbb{R}$

  Gradient vector: $\nabla f(x) = \left[\frac{\partial}{\partial x} f(x)\right]^\top \in \mathbb{R}^n$

Gradient descent

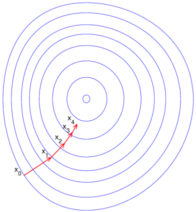
- Problem: $\min_{x\in\mathbb{R}^n} f(x)$ for smooth objective function: $f : \mathbb{R}^n \to \mathbb{R}$

  Gradient vector: $\nabla f(x) = \left[\frac{\partial}{\partial x} f(x)\right]^\top \in \mathbb{R}^n$

- Plain gradient descent: iterative steps in the direction $-\nabla f(x)$:

Input: initial $x \in \mathbb{R}^n$, function $\nabla f(x)$, stepsize α, tolerance θ
Output: x

repeat
    $x \leftarrow x - \alpha \nabla f(x)$
until $|\Delta x| < \theta$ [perhaps for 10 iterations in sequence]



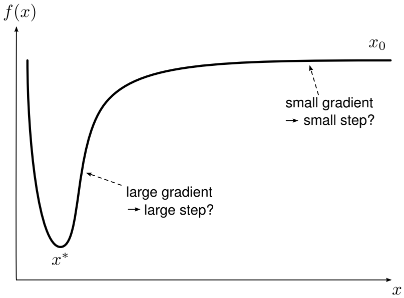
- Plain gradient descent may not be efficient
- Two core issues (for any downhill method):

1. Stepsize

2. Step direction

Stepsize

- Making steps proportional to $\nabla f(x)$?



- We need methods that robustly adapt stepsize

Stepsize Adaptation: Backtracking Line Search

Input: initial $x \in \mathbb{R}^n$, functions $f(x)$ and $\nabla f(x)$, tolerance θ, parameters (defaults: $\varrho_\alpha^+ = 1.2, \varrho_\alpha^- = 0.5, \delta_{\text{max}} = \infty, \varrho_{\text{ls}} = 0.01$)

1: initialize stepsize $\alpha = 1$
2: **repeat**
3: $\quad \delta \leftarrow -\frac{\nabla f(x)}{|\nabla f(x)|}$ *// (alternative: $\delta = -\nabla f(x)$)*
4: $\quad$ **while** $f(x + \alpha\delta) > f(x) + \varrho_{\text{ls}} \nabla f(x)^\top (\alpha\delta)$ **do** ***// line search***
5: $\quad\quad \alpha \leftarrow \varrho_\alpha^- \alpha$ *// REJECT & decrease stepsize*
6: $\quad$ **end while**
7: $\quad x \leftarrow x + \alpha\delta$ *// ACCEPT*
8: $\quad \alpha \leftarrow \min\{\varrho_\alpha^+ \alpha, \delta_{\text{max}}\}$ *// increase stepsize*
9: **until** $|\alpha\delta| < \theta$ *// perhaps for 10 iterations in sequence*

- α determines the absolute stepsize
- Guaranteed monotonicity (by construction)
  (“Typically” ensures convergence to locally convex minima; see later)

Backtracking line search

- Line search in general denotes the problem

$$\min_{\alpha \geq 0} f(x + \alpha\delta)$$

  for some step direction δ.

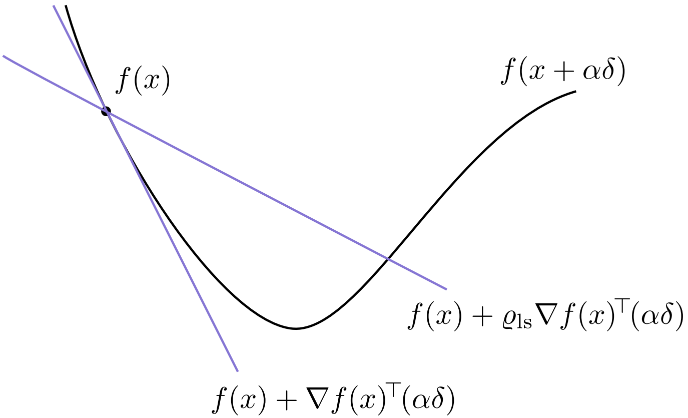
- The most common line search is **backtracking**, which decreases α as long as

$$f(x + \alpha\delta) > f(x) + \varrho_{\text{ls}} \nabla f(x)^\top (\alpha\delta)$$

  ϱ_α^- describes the stepsize decrement in case of a rejected step
  ϱ_{ls} describes a minimum desired decrease in $f(x)$

- Boyd at al: typically $\varrho_{\text{ls}} \in [0.01, 0.3]$ and $\varrho_\alpha^- \in [0.1, 0.8]$

Backtracking line search



Wolfe Conditions

- The 1st Wolfe condition ("sufficient decrease condition")

$$f(x + \alpha\delta) \le f(x) + \varrho_{\text{ls}} \nabla f(x)^\top (\alpha\delta)$$

  requires a decrease of f at least ϱ_{ls}-times "as expected"

- The 2nd (stronger) Wolfe condition ("curvature condition")

$$|\nabla f(x + \alpha\delta)^\top \delta| \le \varrho_{\text{ls2}} |\nabla f(x)^\top \delta|$$

  requires a decrease of the slope by a factor ϱ_{ls2}.
  $\varrho_{\text{ls2}} \in (\varrho_{\text{ls}}, \frac{1}{2})$ (for conjugate gradient)

- See Nocedal et al., Section 3.1 & 3.2 for more general proofs of convergence of any method that ensures the Wolfe conditions after each line search

Convergence for strongly convex functions

- **Theorem** (Exponential convergence on convex functions)
  - Let $f : \mathbb{R}^n \to \mathbb{R}$ be an objective function
  - with eigenvalues λ of the Hessian $\nabla^2 f(x)$ bounded by $m < \lambda < M$, with $m > 0$, $\forall x \in \mathbb{R}^n$
  - Then gradient descent with backtracking line search converges exponentially with convergence rate $(1 - 2\frac{m}{M}\varrho_{\text{ls}}\varrho_\alpha^-)$.

    More precisely: Let x_i and x_{i+1} be two accepted iterates (backtracking line search started at x_i and stopped by accepting x_{i+1}), then

$$f(x_{i+1}) - f_{\text{Min}} \leq \left[1 - \frac{2m\varrho_{\text{ls}}\varrho_\alpha^-}{M}\right] (f(x_i) - f_{\text{Min}}) \ .$$

(I leave the proof to the exercises.)

Discussion of Complexity

- Each line search reduces $f(x)$ at least by

$$f(x_{\text{new}}) - f_{\text{Min}} \;\le\; \Big[1 - \frac{2m\varrho_{\text{ls}}\varrho_\alpha^-}{M}\Big]\,(f(x_{\text{old}}) - f_{\text{Min}})$$

Discussion of Complexity

- Each line search reduces $f(x)$ at least by

$$f(x_{\text{new}}) - f_{\text{Min}} \;\le\; \left[1 - \frac{2m\varrho_{\text{ls}}\varrho_\alpha^-}{M}\right] (f(x_{\text{old}}) - f_{\text{Min}})$$

- How does it scale with the decision space dimension n?

Discussion of Complexity

- Each line search reduces $f(x)$ at least by

$$f(x_{\text{new}}) - f_{\text{Min}} \;\le\; \Big[1 - \frac{2m\varrho_{\text{ls}}\varrho_\alpha^-}{M}\Big]\,(f(x_{\text{old}}) - f_{\text{Min}})$$

- How does it scale with the decision space dimension n?

- What's the intuition behind it being independent of n?