Unconstrained minimization: gradient descent

15.093: Optimization

Dimitris Bertsimas
Alexandre Jacquillat

Sloan School of Management
Massachusetts Institute of Technology

Unconstrained minimization

Formulation (Unconstrained minimization)

Let $f\ \mathbb{R}^n \mapsto \mathbb{R}$ be a continuous (usually differentiable) function. The associated unconstrained non-linear optimization problem is given by

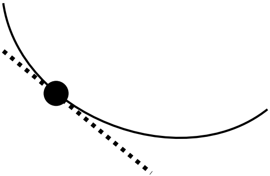
$$\min_{\boldsymbol{x} \in \mathbb{R}^n} \ f(\boldsymbol{x})$$

- Notation: $z^* = \min f(\boldsymbol{x})$ and $\boldsymbol{x}^* \in \arg\min f(\boldsymbol{x})$
- Focus on the difficulties at the core of non-linear optimization
  - The techniques used for unconstrained non-linear minimization also lie at the core of algorithms for constrained non-linear minimization
- Sample applications of unconstrained non-linear minimization
  - Machine learning: linear regression, logistic regression, neural networks
  - Signal processing: interpolation, extrapolation, de-noising
  - Robotics: position, orientation, inverse kinematics

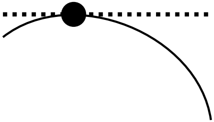
Necessary and sufficient optimality conditions

Some intuition in one dimension

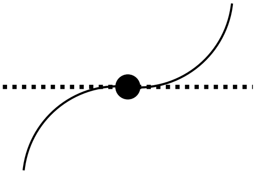
Case with $f'(x) \neq 0$



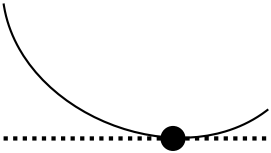
Case with $f'(x) = 0$ and $f''(x) < 0$



Case with $f'(x) = 0$ and $f''(x) = 0$



Case with $f'(x) = 0$ and $f''(x) > 0$



- Necessity of first-order condition at a local optimum: $\nabla f(\boldsymbol{x}^*) = 0$
- Necessity of additional condition to guarantee local optimality, e.g., second-order condition convexity

Necessary optimality conditions

Theorem

Let $f : \mathbb{R}^n \mapsto \mathbb{R}$ be a twice continuously differentiable function.
If $\boldsymbol{x}^* \in \mathbb{R}^n$ is a local minimum of $f(\cdot)$, then

$$\nabla f(\boldsymbol{x}^*) = \mathbf{0} \text{ and } \nabla^2 f(\boldsymbol{x}^*) \succeq \mathbf{0}$$

- Proof of $\nabla f(\boldsymbol{x}^*) = \mathbf{0}$ (first-order necessary condition)
  - For all $\boldsymbol{d} \in \mathbb{R}^n, \alpha > 0$, $\frac{f(\boldsymbol{x}^*+\alpha \boldsymbol{d}) - f(\boldsymbol{x}^*)}{\alpha} \geq 0$
  - Take limits as $\alpha \to 0$: for all $\boldsymbol{d} \in \mathbb{R}^n$: $\nabla f(\boldsymbol{x}^*)^\top \boldsymbol{d} \geq 0$
  - Apply the inequality to $\boldsymbol{e}_i \in \mathbb{R}^n$ and $-\boldsymbol{e}_i \in \mathbb{R}^n$: $\nabla_i f(\boldsymbol{x}^*) = 0$ for all i

- Proof of $\nabla^2 f(\boldsymbol{x}^*) \succeq \mathbf{0}$ (second-order necessary condition)
  - Taylor series expansion: $\rho(\cdot)$ with $\rho(\boldsymbol{y}) \to 0$ as $\boldsymbol{y} \to \mathbf{0}$, such that

$$\underbrace{f(\boldsymbol{x}^* + \alpha \boldsymbol{d}) - f(\boldsymbol{x}^*)}_{\geq 0} = \alpha \underbrace{\nabla f(\boldsymbol{x}^*)^\top}_{=\mathbf{0}} \boldsymbol{d} + \frac{\alpha^2}{2} \boldsymbol{d}^\top \nabla^2 f(\boldsymbol{x}^*) \boldsymbol{d} + \alpha^2 \|\boldsymbol{d}\|^2 \rho(\alpha \boldsymbol{d})$$

$$\alpha \to 0 : \implies \boldsymbol{d}^\top \nabla^2 f(\boldsymbol{x}^*) \boldsymbol{d} \geq 0$$

Sufficient optimality conditions

Definition (Positive definite matrix)

A symmetric matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ is positive definite, written $\boldsymbol{A} \succ \boldsymbol{0}$, if $\boldsymbol{u}^\top \boldsymbol{A} \boldsymbol{u} > 0$ for all $\boldsymbol{u} \neq \boldsymbol{0}$.
Equivalently, $\boldsymbol{A} \succ \boldsymbol{0}$ if and only if all its eigenvalues are positive.

Theorem

Let $f : \mathbb{R}^n \mapsto \mathbb{R}$ be twice continuously differentiable.
If $\nabla f(\boldsymbol{x}^*) = \boldsymbol{0}$ and $\nabla^2 f(\boldsymbol{x}) \succ \boldsymbol{0}$, then $\boldsymbol{x}^*$ is a local minimum.

- Proof:
  - Let $\lambda > 0$ be the smallest eigenvalue of $\nabla^2 f(\boldsymbol{x})$
  - Taylor series expansion: $\rho(\cdot)$ with $\rho(\boldsymbol{y}) \to 0$ as $\boldsymbol{y} \to \boldsymbol{0}$, such that

$$
\begin{aligned}
f(\boldsymbol{x}^* + \boldsymbol{y}) - f(\boldsymbol{x}^*) &= \underbrace{\nabla f(\boldsymbol{x}^*)}_{=\boldsymbol{0}}^\top \boldsymbol{y} + \frac{1}{2} \underbrace{\boldsymbol{y}^\top \nabla^2 f(\boldsymbol{x}^*) \boldsymbol{y}}_{\geq \lambda \|\boldsymbol{y}\|^2} + \|\boldsymbol{y}\|^2 \rho(\boldsymbol{y}) \\
&\geq \frac{\lambda}{2} \|\boldsymbol{y}\|^2 + \|\boldsymbol{y}\|^2 \rho(\boldsymbol{y}) \\
&\geq 0, \text{ for } \|\boldsymbol{y}\| \text{ small enough.}
\end{aligned}
$$

Example

- Consider the following two-dimensional function:

$$f(\boldsymbol{x}) = \frac{1}{2}x_1^2 + x_1.x_2 + 2x_2^2 - 4x_1 - 4x_2 - x_2^3$$

- Compute its gradient and its Hessian:

$$\nabla f(\boldsymbol{x}) = (x_1 + x_2 - 4,\ x_1 + 4x_2 - 4 - 3x_2^2)^\top$$

$$\nabla^2 f(\boldsymbol{x}) = \begin{bmatrix} 1 & 1 \\ 1 & 4 - 6x_2 \end{bmatrix}$$

→ Two candidates for local minima: $\boldsymbol{x}^* = (4, 0)$ and $\overline{\boldsymbol{x}} = (3, 1)$

- $\overline{\boldsymbol{x}}$ is not a local minimum from second-order information

$$\nabla^2 f(\overline{\boldsymbol{x}}) = \begin{bmatrix} 1 & 1 \\ 1 & -2 \end{bmatrix} \text{ is an indefinite matrix}$$

- $\boldsymbol{x}^*$ is a local minimum:

$$\nabla^2 f(\boldsymbol{x}^*) = \begin{bmatrix} 1 & 1 \\ 1 & 4 \end{bmatrix} \succ \mathbf{0}$$

Another example

- Consider the following two-dimensional function:

$$f(\boldsymbol{x}) = x_1^3 + x_2^2$$

- Compute its gradient and its Hessian:

$$\nabla f(\boldsymbol{x}) = (3x_1^2, 2x_2)^\top$$

$$\nabla^2 f(\boldsymbol{x}) = \begin{bmatrix} 6x_1 & 0 \\ 0 & 2 \end{bmatrix}$$

→ One candidate for local minimum: $\boldsymbol{x}^* = (0, 0)$

- $\boldsymbol{x}^*$ is still a candidate a local minimum from second-order information

$$\nabla^2 f(\boldsymbol{x}^*) = \begin{bmatrix} 0 & 0 \\ 0 & 2 \end{bmatrix} \succeq \mathbf{0}$$

- However, $\boldsymbol{x}^*$ is not a local minimum.

$$\widetilde{\boldsymbol{x}} = (-\varepsilon, 0)^\top \implies f(\widetilde{\boldsymbol{x}}) = -\varepsilon^3 < 0 = f(\boldsymbol{x}^*)$$

The case of convexity

Proposition (Reminder: first-order convexity condition)

Let $f : \mathbb{R}^n \mapsto \mathbb{R}$ be a differentiable function. It is convex if and only:

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^\top (\boldsymbol{y} - \boldsymbol{x}), \ \forall \boldsymbol{x}, \boldsymbol{y}$$

Proposition (Reminder: second-order condition)

Let $f\cdot)$ be a twice continuously differentiable function.
f is a convex function if and only if $H(\boldsymbol{x}) \succeq \mathbf{0}$ for all $\boldsymbol{x} \in \mathbb{R}^n$

Theorem

Let $f(\boldsymbol{x})$ be a continuously differentiable convex function.
Then $\boldsymbol{x}^*$ is a global minimum of f <u>if and only if</u> $\nabla f(\boldsymbol{x}^*) = \mathbf{0}$.

- Proof: If f is convex and $\nabla f(\boldsymbol{x}^*) = \mathbf{0}$, we have:

$$f(\boldsymbol{x}) - f(\boldsymbol{x}^*) \geq \nabla f(\boldsymbol{x}^*)^\top (\boldsymbol{x} - \boldsymbol{x}^*) = 0$$

Example: fitting a linear regression model

- Linear regression model with ℓ_2-regularization: $\min f(\boldsymbol{\beta})$ with

$$f(\boldsymbol{\beta}) = \frac{1}{2}\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{m} x_{ij}\beta_j\right)^2 + \frac{\lambda}{2}\sum_{j=1}^{m}\beta_j^2 = \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \frac{\lambda}{2}\|\boldsymbol{\beta}\|^2$$

- Differentiation of the loss function:

$$\nabla f(\boldsymbol{\beta}) = -\boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) + \lambda\boldsymbol{\beta}$$
$$\nabla^2 f(\boldsymbol{\beta}) = \boldsymbol{X}^\top\boldsymbol{X} + \lambda\boldsymbol{I}$$

- Importance of ℓ_2-regularization in case $\boldsymbol{X}$ does not have full rank
  - $\boldsymbol{X}^\top\boldsymbol{X}$ is positive semidefinite: $\boldsymbol{u}^\top\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{u} = \|\boldsymbol{X}\boldsymbol{u}\|^2 \geq 0, \forall \boldsymbol{u} \in \mathbb{R}^n$
  - $\boldsymbol{X}^\top\boldsymbol{X} + \lambda\boldsymbol{I}$ is positive definite if $\lambda > 0$: $\boldsymbol{u}^\top(\boldsymbol{X}^\top\boldsymbol{X} + \lambda\boldsymbol{I})\boldsymbol{u} = \|\boldsymbol{X}\boldsymbol{u}\|^2 + \lambda\|\boldsymbol{u}\|^2 > 0, \forall \boldsymbol{u} \neq \boldsymbol{0}$

→ A convex quadratic optimization problem: $\nabla^2 f(\boldsymbol{\beta}) \succeq \boldsymbol{0}$

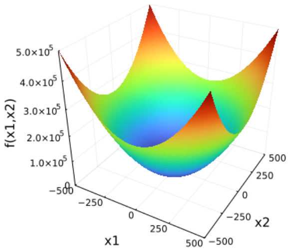
→ The problem admits a minimum at a stationary point

$$\boldsymbol{\beta}^* = (\boldsymbol{X}^\top\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}^\top\boldsymbol{y}$$

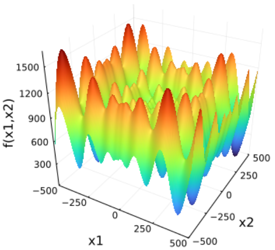
Summary and implications

- A local minimum can only occur in a stationary point: $\nabla f(\boldsymbol{x}^*) = \mathbf{0}$
- Convex optimization: any stationary point is a global minimum!
- Otherwise, local optimality depends on second-order conditions
  - If $\nabla^2 f(\boldsymbol{x}^*) \succ 0$, then $\boldsymbol{x}^*$ is a local minimum
  - If $\nabla^2 f(\boldsymbol{x}^*) \succeq 0$, then $\boldsymbol{x}^*$ *can be* a local minimum
  - If $\nabla^2 f(\boldsymbol{x}^*)$ is indefinite, then $\boldsymbol{x}^*$ is not a local minimum
- In non-convex optimization, identifying global minima is very hard

Convex optimization



Non-convex optimization



Algorithms for unconstrained optimization

A generic descent method

- Seek a stationary point as a candidate for local minimum

$$\nabla f(\boldsymbol{x}) = \mathbf{0}$$

Algorithm

1. *Initialization: starting point* $\boldsymbol{x}^0 \in \mathbb{R}^n$*, and iteration counter* $k = 0$
2. *Repeat, until termination criterion is reached (e.g.,* $\|\nabla f(\boldsymbol{x}^k)\| \leq \eta$*)*
   - 2.1 *Update iteration counter:* $k \leftarrow k + 1$
   - 2.2 *Determine a descent direction* $\boldsymbol{d}^k$*, such that* $\nabla f(\boldsymbol{x}^k)^\top \boldsymbol{d}^k < 0$
   - 2.3 *Determine a step size* $\alpha^k > 0$
   - 2.4 *Update* $\boldsymbol{x}^{k+1} \leftarrow \boldsymbol{x}^k + \alpha^k \boldsymbol{d}^k$

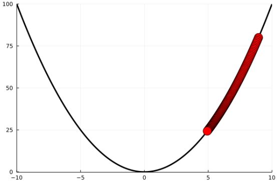
- A descent algorithm: improvement in the solution at each iteration as long as the step size is small enough and $\nabla f(\boldsymbol{x}^k)^\top \boldsymbol{d}^k < 0$

$$f(\boldsymbol{x}^{k+1}) \approx f(\boldsymbol{x}^k) + \alpha^k \nabla f(\boldsymbol{x}^k)^\top \boldsymbol{d} < f(\boldsymbol{x}^k), \ \forall k = 0, 1, \cdots$$

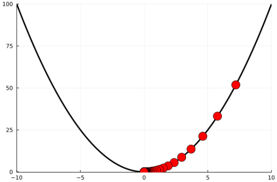
Impact of step size on algorithm convergence

- Core trade-off in setting step sizes
  - Small step sizes: slow convergence, with little progress per iteration
  - Large step sizes: unstable behavior
  - In-between, faster convergence can be achieved
- The appropriate step size depends on the shape of the function to minimize, the descent direction, and the progress of the algorithm

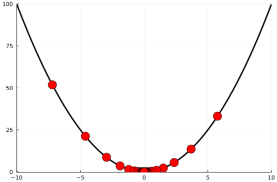
Small step size



Intermediate step size



Large step size



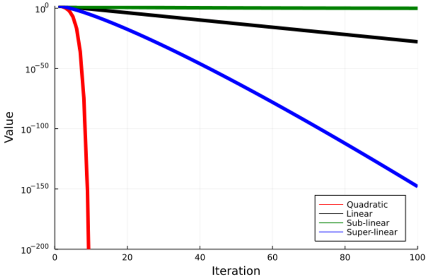
Rate of convergence of algorithms

Definition

A convergence sequence $z_1, \cdots, z_n \to z$ has order of convergence p and rate of convergence β if

$$\lim_{k\to\infty} \frac{|z_{k+1} - z|}{|z_k - z|^p} = \beta$$

- Linear convergence: $p^* = 1$, $0 < \beta < 1$, e.g., $z_k = a^k$, $a < 1$
- Sub-linear convergence: $p^* = 1$, $\beta = 1$, e.g., $z_k = 1/k$
- Super-linear convergence: $p^* = 1$, $\beta = 0$, e.g., $z_k = (1/k)^k$
- Quadratic convergence: $p^* = 2$, e.g., $z_k = a^{2^k}$, $a < 1$



Gradient descent: convergence analysis

Gradient descent algorithm

Algorithm

1. *Initialization: starting point* $\boldsymbol{x}^0 \in \mathbb{R}^n$*, and iteration counter* $k = 0$
2. *Repeat, until stopping criterion is reached*
   - 2.1 *Update iteration counter:* $k \leftarrow k + 1$
   - 2.2 *Choose descent direction* $\boldsymbol{d}^k = -\nabla f(\boldsymbol{x}^k)$
   - 2.3 *Determine a step size* $\alpha^k > 0$
   - 2.4 *Update* $\boldsymbol{x}^{k+1} \leftarrow \boldsymbol{x}^k + \alpha^k \boldsymbol{d}^k$

- Motivation: the gradient rule defines a valid descent direction

$$f(\boldsymbol{x}^{k+1}) \approx f(\boldsymbol{x}^k) - \alpha \|\nabla f(\boldsymbol{x}^k)\|^2 < f(\boldsymbol{x}^k)$$

$\rightarrow$ Gradient descent defines a monotonically decreasing sequence

$$f(\boldsymbol{x}^0) > f(\boldsymbol{x}^1) > \cdots > f(\boldsymbol{x}^k) > \cdots$$

- Under assumptions, $\boldsymbol{x}^0, \boldsymbol{x}^1, \cdots$, will converge to a stationary point

Further assumptions: smoothness and strong convexity

Definition (smoothness)

A function $f(\cdot)$ is M-smooth if its gradient is M-Lipschitz continuous:

$$\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\| \leq M\|\boldsymbol{x} - \boldsymbol{y}\|, \ \forall \boldsymbol{x}, \boldsymbol{y}$$

Definition (strong convexity)

A function $f(\cdot)$ is m-strongly convex if:

$$f(\boldsymbol{x}) - \frac{m}{2}\|\boldsymbol{x}\|^2 \text{ is convex.}$$

Proposition

If f is M-smooth, it satisfies:

$$f(\boldsymbol{y}) \leq f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^\top (\boldsymbol{y} - \boldsymbol{x}) + \frac{M}{2}\|\boldsymbol{y} - \boldsymbol{x}\|^2, \ \forall \boldsymbol{x}, \boldsymbol{y}$$

If f is m-strongly convex and continuously differentiable, it satisfies

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^\top (\boldsymbol{y} - \boldsymbol{x}) + \frac{m}{2}\|\boldsymbol{y} - \boldsymbol{x}\|^2, \ \forall \boldsymbol{x}, \boldsymbol{y}$$

Gradient descent with constant step size: convergence

Theorem (Convergence with constant step size)

Assume that f is M-smooth and convex. The gradient descent algorithm with fixed step size $\alpha \leq 1/M$ converges in $\mathcal{O}(1/k)$ to a global minimum:

$$f(\boldsymbol{x}^k) - z^* \leq \frac{1}{2\alpha k}\|\boldsymbol{x}^0 - \boldsymbol{x}^*\|^2,\ \forall k \geq 0$$

- Sub-linear convergence: $f(\boldsymbol{x}^k) - z^* \leq \frac{1}{2\alpha k}\|\boldsymbol{x}^0 - \boldsymbol{x}^*\|^2$
- Solution within ε of the optimum after at most

$$\frac{\|\boldsymbol{x}^0 - \boldsymbol{x}^*\|^2}{2\alpha\varepsilon} \text{ iterations}$$

$\rightarrow$ Drivers of algorithm performance

- Initial condition: the weaker $\boldsymbol{x}^0$, the slower the convergence
- Tolerance: the smaller ε, the slower the convergence
- Step size: convergence is faster when α is larger, yet not too large

Gradient descent with constant step size: proof

- Using $\boldsymbol{x}^{k+1} = \boldsymbol{x}^k - \alpha \nabla f(\boldsymbol{x}^k)$ and M-smoothness, we have:

$$f(\boldsymbol{x}^{k+1}) \leq f(\boldsymbol{x}^k) + \nabla f(\boldsymbol{x}^k)^\top \underbrace{(\boldsymbol{x}^{k+1} - \boldsymbol{x}^k)}_{-\alpha \nabla f(\boldsymbol{x}^k)} + \underbrace{M/2}_{\leq 1/(2\alpha)} \cdot \underbrace{\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^k\|^2}_{=\alpha^2 \|\nabla f(\boldsymbol{x}^k)\|^2}$$

$$\implies f(\boldsymbol{x}^{k+1}) \leq f(\boldsymbol{x}^k) - \frac{\alpha}{2}\|\nabla f(\boldsymbol{x}^k)\|^2 = f(\boldsymbol{x}^k) - \frac{1}{2\alpha}\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^k\|^2$$

- By convexity: $f(\boldsymbol{x}^*) \geq f(\boldsymbol{x}^k) + \nabla f(\boldsymbol{x}^k)^\top (\boldsymbol{x}^* - \boldsymbol{x}^k)$. We obtain

$$\begin{aligned} f(\boldsymbol{x}^{k+1}) - f(\boldsymbol{x}^*) &\leq -\nabla f(\boldsymbol{x}^k)^\top (\boldsymbol{x}^* - \boldsymbol{x}^k) - \frac{1}{2\alpha}\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^k\|^2 \\ &= \cdots \text{ [after some algebra]} \\ &= \frac{1}{2\alpha}\left(\|\boldsymbol{x}^k - \boldsymbol{x}^*\|^2 - \|\boldsymbol{x}^{k+1} - \boldsymbol{x}^*\|^2\right) \end{aligned}$$

- We conclude by telescoping the sum and exploiting monotonicity:

$$\underbrace{\sum_{i=1}^k (f(\boldsymbol{x}^k) - f(\boldsymbol{x}^*))}_{=k(f(\boldsymbol{x}^k) - f(\boldsymbol{x}^*))} \leq \sum_{i=1}^k (f(\boldsymbol{x}^i) - f(\boldsymbol{x}^*)) \leq \underbrace{\frac{1}{2\alpha}\left(\|\boldsymbol{x}^0 - \boldsymbol{x}^*\|^2 - \|\boldsymbol{x}^k - \boldsymbol{x}^*\|^2\right)}_{\leq \frac{1}{2\alpha}\|\boldsymbol{x}^0 - \boldsymbol{x}^*\|^2}$$

Gradient descent with exact line search: convergence

- Gradient descent with exact line search: at each iteration k, choose step size α^k to maximize the one-step improvement:

$$\alpha^k \in \arg\min \ f(\boldsymbol{x}^k + \alpha^k \boldsymbol{d}^k)$$

Theorem (Convergence with exact line search)

Assume that f is M-smooth and m-strongly convex. The gradient descent algorithm with exact line search converges in $\mathcal{O}(c^k)$, with $c = 1 - m/M$

$$f(\boldsymbol{x}^k) - z^* \leq c^k(f(\boldsymbol{x}^{(0)} - z^*), \ \forall k \geq 0$$

- Stronger convergence with f strongly convex and optimized step sizes
  - Linear convergence: $f(\boldsymbol{x}^k) - z^* \leq c^k(f(\boldsymbol{x}^{(0)} - z^*)$
  - Solution within ε of the optimum after at most

$$\frac{\log(f(\boldsymbol{x}^{(0)} - z^*) - \log(\varepsilon)}{\log(1/c)} \text{ iterations}$$

Gradient descent with exact line search: proof

- Due to the M-smoothness of f:

$$f(\boldsymbol{x}^k - \alpha^k \nabla f(\boldsymbol{x}^k)) \leq f(\boldsymbol{x}^k) - \alpha^k \|\nabla f(\boldsymbol{x}^k)\|^2 + \frac{M(\alpha^k)^2}{2} \|\nabla f(\boldsymbol{x}^k)\|^2$$

- By optimizing over α^k and subtracting z^* on both sides, we obtain:

$$(f(\boldsymbol{x}^{k+1}) - z^*) \leq (f(\boldsymbol{x}^k) - z^*) - \frac{1}{2M} \|\nabla f(\boldsymbol{x}^k)\|^2$$

- Due to the strong convexity of f, we have

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}^k) + \nabla f(\boldsymbol{x}^k)^\top (\boldsymbol{y} - \boldsymbol{x}^k) + \frac{m}{2} \|\boldsymbol{y} - \boldsymbol{x}^k\|^2, \ \forall \boldsymbol{y}$$

- By minimizing over y on both sides, we obtain:

$$z^* \geq f(\boldsymbol{x}^k) - \frac{1}{2m} \|\nabla f(\boldsymbol{x}^k)\|^2$$

- We conclude:

$$(f(\boldsymbol{x}^{k+1}) - z^*) \leq \left(1 - \frac{m}{M}\right) (f(\boldsymbol{x}^k) - z^*)$$

$$\implies (f(\boldsymbol{x}^k) - z^*) \leq \left(1 - \frac{m}{M}\right)^k (f(\boldsymbol{x}^{(0)}) - z^*)$$

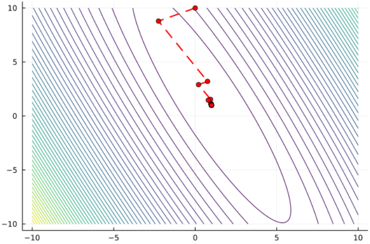
Example

$$\min \; f(x_1, x_2) = 5x_1^2 + x_2^2 + 4x_1x_2 - 14x_1 - 6x_2 + 20$$

$$\begin{pmatrix} d_1^k \\ d_2^k \end{pmatrix} = -\nabla f(x_1^k, x_2^k) = \begin{pmatrix} -10x_1^k - 4x_2^k + 14 \\ -2x_2^k - 4x_1^k + 6 \end{pmatrix}$$

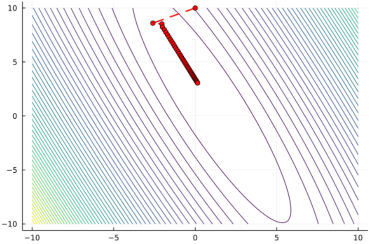
Exact line search

$$\alpha^k = \frac{(d_1^k)^2 + (d_2^k)^2}{2(5(d_1^k)^2 + (d_2^k)^2 + 4d_1^k d_2^k)}$$



Constant step size

$$\alpha^k = 0.1$$



Advanced topics

Role of the condition number in quadratic optimization

Definition (condition number)

The condition number of a non-singular matrix $\boldsymbol{Q}$ is the ratio of its largest-to-smallest eigenvalues: $\kappa(\boldsymbol{Q}) = \frac{\lambda_{\max}}{\lambda_{\min}} \geq 1$

Theorem (Kantorovich, Nobel Prize in Economics, 1975)

If $f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^\top \boldsymbol{Q}\boldsymbol{x} - \boldsymbol{c}^\top \boldsymbol{x} + b$ *and* $\boldsymbol{Q} \succ 0$, *then*

$$(f(\boldsymbol{x}^{k+1}) - z^*) \leq \left(\frac{\kappa(\boldsymbol{Q}) - 1}{\kappa(\boldsymbol{Q}) + 1}\right)^2 (f(\boldsymbol{x}^k) - z^*)$$

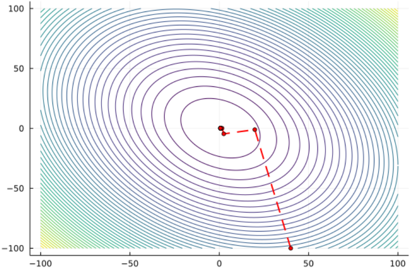
- The condition number plays a critical role in the convergence of gradient descent algorithms for quadratic functions
  - $\kappa(\boldsymbol{Q}) \approx 1$: "well-conditioned" matrix, fast convergence
  - $\kappa(\boldsymbol{Q}) >> 1$: "ill-conditioned" matrix, slow convergence

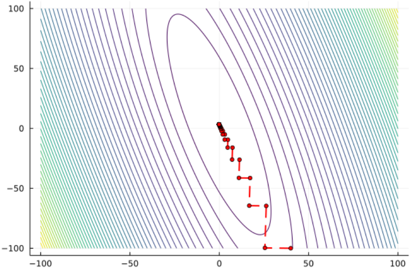
Role of the condition number: illustration

$$\min \ f(x_1, x_2) = \frac{1}{2}\boldsymbol{x}^\top \boldsymbol{Q}\boldsymbol{x} - \boldsymbol{c}^\top \boldsymbol{x} + 10 \qquad \boldsymbol{d}^k = -\nabla f(\boldsymbol{x}^k) = -\boldsymbol{Q}\boldsymbol{x}^k + \boldsymbol{c}$$

$$\text{Exact line search: } \alpha^k = \frac{(\boldsymbol{d}^k)^\top \boldsymbol{d}^k}{(\boldsymbol{d}^k)\boldsymbol{Q}\boldsymbol{d}^k}$$

$$\boldsymbol{Q} = \begin{bmatrix} 20 & 5 \\ 5 & 16 \end{bmatrix} \to \kappa(\boldsymbol{Q}) = 1.85 \qquad \boldsymbol{Q} = \begin{bmatrix} 20 & 5 \\ 5 & 2 \end{bmatrix} \to \kappa(\boldsymbol{Q}) = 30.23$$





Beyond quadratic optimization

- Recall the drivers of the performance of the gradient descent algorithm
  - Initial condition: the weaker $\boldsymbol{x}^0$, the slower the convergence
  - Tolerance: the smaller ε, the slower the convergence
  - Function f: the higher the constant c, the slower the convergence
- The constant $c = 1 - \frac{m}{M}$ relates to the convex shape of the function f
- In fact, c depends on the condition number of the Hessian of f around the optimal value of the optimization problem

$$\kappa(\nabla^2 f(\boldsymbol{x}^*)) \lesssim \frac{M}{m}$$

$\rightarrow$ Beyond quadratic optimization: the convergence rate is impacted by the condition number of the Hessian matrix

  - Strong convergence when the Hessian is well-conditioned
  - Slow convergence when the Hessian is ill-conditioned

Extension: norm-based steepest gradient descent

Definition (steepest descent)

For any norm $\|\cdot\|$, the steepest descent direction is given as follows:

$$\boldsymbol{x}^{k+1} = \boldsymbol{x}^k - \alpha^k \boldsymbol{d}^k, \text{ with } \boldsymbol{d}^k \in \arg\min\{\nabla f(\boldsymbol{x}^k)\boldsymbol{v} : \|\boldsymbol{v}\| = 1\}$$

- Motivation: finding the direction with the strongest improvement

$$f(\boldsymbol{x}+\boldsymbol{v}) \approx f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^\top \boldsymbol{v}$$

- ℓ_2-norm $\|\cdot\|_2 \implies \boldsymbol{d}^k = -\nabla f(\boldsymbol{x}^k)$
  - Cauchy–Schwarz inequality: $|\nabla f(\boldsymbol{x}^k)\boldsymbol{v}| \leq \|\nabla f(\boldsymbol{x}^k)\|\|\boldsymbol{v}\| = \|\nabla f(\boldsymbol{x}^k)\|$

  → Gradient descent is merely steepest descent with the ℓ_2-norm
- Quadratic norm $\|\boldsymbol{z}\|_{\boldsymbol{P}} = \sqrt{\boldsymbol{z}^\top \boldsymbol{P} \boldsymbol{z}} \implies \boldsymbol{d}^k = -\boldsymbol{P}^{-1}\nabla f \boldsymbol{x}$

  → Equivalent to gradient descent with re-scaling $\overline{\boldsymbol{x}} = \boldsymbol{P}^{1/2}\boldsymbol{x}$ to circumvent ill-conditioned problems
- ℓ_1-norm $\|\cdot\|_1 \implies \boldsymbol{d}^k = -\frac{\partial f}{\partial x_i}\boldsymbol{e}_i$

  → Equivalent to coordinate descent: one variable at a time

Conclusion

Summary

Takeaway

Unconstrained minimization is the core problem in non-linear optimization, with applications in machine learning, signal processing, robotics, etc.

Takeaway

Descent methods can find local optima. They imply global optimality if the function is convex; global optimization is much harder otherwise.

Takeaway

Gradient descent is a simple algorithm that exhibits sub-linear convergence with constant step sizes and linear convergence with optimized step sizes.

Takeaway

Convergence is greatly impacted by the condition number of $\nabla^2 f(\boldsymbol{x}^*)$.