MA427 Lecture 10
Gradient descent

Giacomo Zambelli

Department of Mathematics

THE LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE

18 March, 2019

Today's lecture

- The gradient descent method
- Conditioning the function
- Convergence analysis of gradient descent
- Constrained convex optimisation

Descent methods

Unconstrained optimisation problem:
Convex function $f : \mathbb{R}^n \to \mathbb{R}$:

$$\min_{x \in \textbf{dom}\, f} f(x).$$

Descent method:

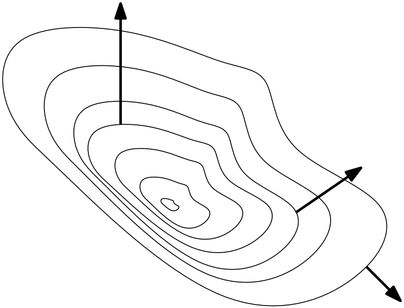
- Start from *initial point* $x^{(0)} \in \textbf{dom}\, f$.
- Construct a series of points $x^{(1)}, x^{(2)}, \ldots$ with $f(x^{(k)}) > f(x^{(k+1)})$.
- Update:
$$x^{(k+1)} = x^{(k)} + t_k \Delta x^{(k)}.$$
- *Search direction:* $\Delta x^{(k)}$.
- *Step size:* t_k.

Gradient descent

- Let $f : \mathbb{R}^n \to \mathbb{R}$ be differentiable
- x^* is a global minimum point if and only if $\nabla f(x^*) = 0$.
- Taylor expansion:

$$f(x + t\Delta x) \approx f(x) + t\nabla f(x)^\top \Delta x$$

- Natural descent direction: $\Delta x^{(k)} = -\nabla f(x^{(k)})$.



Determining the step size

- *Exact line search*

$$t_k := \text{argmin}_{t \geq 0} f(x^{(k)} - t\nabla f(x^{(k)}))$$

- Minimise the univariate convex function $g(t) = f(x^{(k)} - t\nabla f(x^{(k)}))$ over $t \geq 0$.
- Best possible in this direction, but may be time-consuming to compute.

Determining the step size

- *Exact line search*

$$t_k := \text{argmin}_{t \geq 0} f(x^{(k)} - t\nabla f(x^{(k)}))$$

- Minimise the univariate convex function $g(t) = f(x^{(k)} - t\nabla f(x^{(k)}))$ over $t \geq 0$.
- Best possible in this direction, but may be time-consuming to compute.
- *Constant step size:* fix $t_k = \mu$ throughout.
- Can be difficult to guess the right value.

Backtracking line search

- Two parameters: $0 < \alpha < \frac{1}{2}$, and $0 < \beta < 1$.
- Want to find $t > 0$ with

$$f(x^{(k)} - t\nabla f(x^{(k)})) \leq f(x^{(k)}) - \alpha t \|\nabla f(x)\|^2$$

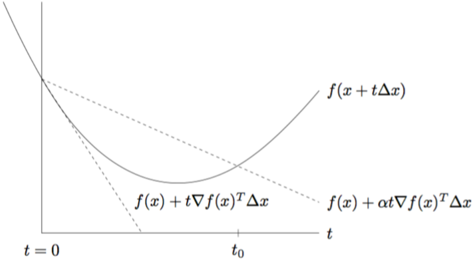
Backtracking line search

- Two parameters: $0 < \alpha < \frac{1}{2}$, and $0 < \beta < 1$.
- Want to find $t > 0$ with

$$f(x^{(k)} - t\nabla f(x^{(k)})) \leq f(x^{(k)}) - \alpha t \|\nabla f(x)\|^2$$

Backtracking line search:

1. Set $t := 1$.
2. *While* $f(x^{(k)} - t\nabla f(x^{(k)})) > f(x^{(k)}) - \alpha t \|\nabla f(x)\|^2$, *update* $t := \beta t$.



Example

- Minimise $f(x_1, x_2) = 4x_1^2 - 4x_1x_2 + 2x_2^2 + 2x_1$ over $\mathbb{R}^2$.
- Optimum $x^* = (-\frac{1}{2}, -\frac{1}{2})$ with value -0.5
- Initial point: $x^{(0)} = (0, 0)$, with value 0, gradient $\nabla f(x^{(0)}) = \binom{2}{0}$.

Example

- Minimise $f(x_1, x_2) = 4x_1^2 - 4x_1x_2 + 2x_2^2 + 2x_1$ over $\mathbb{R}^2$.
- Optimum $x^* = (-\frac{1}{2}, -\frac{1}{2})$ with value -0.5
- Initial point: $x^{(0)} = (0, 0)$, with value 0, gradient $\nabla f(x^{(0)}) = \binom{2}{0}$.
- Exact line search: $t = \frac{1}{8}$, updates to $x^{(1)} = (-\frac{1}{4}, 0)$ with $f(x^{(1)}) = -0.25$.

Example

- Minimise $f(x_1, x_2) = 4x_1^2 - 4x_1x_2 + 2x_2^2 + 2x_1$ over $\mathbb{R}^2$.
- Optimum $x^* = (-\frac{1}{2}, -\frac{1}{2})$ with value -0.5
- Initial point: $x^{(0)} = (0, 0)$, with value 0, gradient $\nabla f(x^{(0)}) = \binom{2}{0}$.
- Exact line search: $t = \frac{1}{8}$, updates to $x^{(1)} = (-\frac{1}{4}, 0)$ with $f(x^{(1)}) = -0.25$.
- Backtracking line search for $\alpha = 0.3$, $\beta = 0.4$.

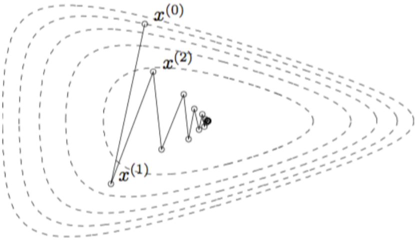
t	$f(-2t, 0)$	$-0.4t$
1	12	-0.4
0.4	0.96	-0.16
0.16	-0.23	-0.064

- $x^{(1)} = (-0.32, 0)$, $f(x^{(1)}) \approx -0.23$.

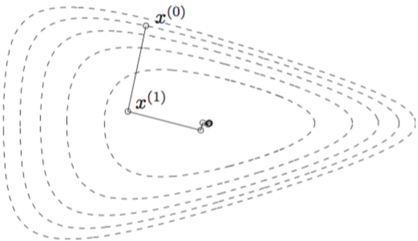
Example

[Boyd & Vandenberghe, Sec 9.3]

$$f(x_1, x_2) = e^{x_1+3x_2-0.1} + e^{x_1-3x_2-0.1} + e^{-x_1-0.1}$$



Backtracking line search,
$\alpha = 0.1$, $\beta = 0.7$



Exact line search

Stopping criterion

- Optimal solution $p^* = f(x^*)$.
  Cannot compute the exact value in most cases!
- Error tolerance: $\varepsilon > 0$.
- $x \in \textbf{dom}\, f$ is a *ε-approximate solution*, if

$$f(x) \geq p^* - \varepsilon.$$

Stopping criterion

- Optimal solution $p^* = f(x^*)$.
  Cannot compute the exact value in most cases!
- Error tolerance: $\varepsilon > 0$.
- $x \in \mathbf{dom}\, f$ is a *ε-approximate solution*, if

$$f(x) \geq p^* - \varepsilon.$$

- Problem: how to decide whether $x^{(k)}$ is already ε-approximate?

Stopping criterion

- Optimal solution $p^* = f(x^*)$.
  Cannot compute the exact value in most cases!
- Error tolerance: $\varepsilon > 0$.
- $x \in \mathbf{dom}\, f$ is a *ε-approximate solution*, if

$$f(x) \geq p^* - \varepsilon.$$

- Problem: how to decide whether $x^{(k)}$ is already ε-approximate?
- *Stopping criterion:* $\|\nabla f(x^{(k)})\| < \delta$ for some threshold $\delta > 0$.
- Can we get a bound $\delta = \delta(\varepsilon)$ such that $\|\nabla f(x^{(k)})\| < \delta$ implies that $x^{(k)}$ is ε-approximate?

Conditioning the function

Notation

Sublevel set of the function

- Initial point $x^{(0)}$:

$$S = \{x \in \mathbf{dom}\, f \;:\; f(x) \leq f(x^{(0)})\}.$$

- We must have $x^* \in S$, and all iterates $x^{(1)}, x^{(2)}, \ldots \in S$.

Notation

Sublevel set of the function

- Initial point $x^{(0)}$:

$$S = \{x \in \mathbf{dom}\, f \; : \; f(x) \leq f(x^{(0)})\}.$$

- We must have $x^* \in S$, and all iterates $x^{(1)}, x^{(2)}, \ldots \in S$.

Ordering positive semidefinite matrices

- $P, Q \in \mathbb{R}^{n \times n}$ positive semidefinite (PSD) matrices.
- $P \preceq Q$: P is *PSD-smaller* than Q, if $Q - P$ is also PSD matrix.
- Equivalently: for any vector $v \in \mathbb{R}^n$,

$$v^\top P v \leq v^\top Q v.$$

Strong convexity

- f is *strongly convex* on S with parameter $m > 0$, if

$$\nabla^2 f(x) \succeq mI_n \quad \text{for every } x \in S.$$

- Equivalently,

$$v^\top \nabla^2 f(x) v \geq m\|v\|^2 \quad \text{for every } x \in S, v \in \mathbb{R}^n.$$

- Also equivalently, all eigenvalues of $\nabla^2 f(x)$ are $\geq m$.

Strong convexity

- f is *strongly convex* on S with parameter $m > 0$, if
$$\nabla^2 f(x) \succeq mI_n \quad \text{for every } x \in S.$$
- Equivalently,
$$v^\top \nabla^2 f(x) v \geq m\|v\|^2 \quad \text{for every } x \in S, v \in \mathbb{R}^n.$$
- Also equivalently, all eigenvalues of $\nabla^2 f(x)$ are $\geq m$.
- *Taylor expansion:* for any $x, y \in S$, we get
$$f(x) \geq f(y) + \nabla f(y)^\top (x - y) + \frac{m}{2}\|x - y\|^2.$$
- Affine functions are not strongly convex!
- A convex quadratic function $f(x) = x^\top Qx + p^\top x$ is strongly convex if and only if Q is positive definite.

Strong convexity

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

Proposition

If f is strongly convex on S, then there exists a unique global minimum point x^*.

Strong convexity

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

Proposition

If f is strongly convex on S, then there exists a unique global minimum point x^*.

Lemma

For $x \in S$, we have

$$f(x) - p^* \leq \frac{1}{2m}\|\nabla f(x)\|^2.$$

In particular, if $\|\nabla f(x)\| \leq (2m\varepsilon)^{1/2}$, then $f(x) - p^* \leq \varepsilon$.

M-smooth functions

- f is *M-Lipschitz smooth*, S for $M > 0$ if

$$\nabla^2 f(x) \preceq M I_n.$$

- Equivalently: all eigenvalues of $\nabla^2 f(x)$ are $\leq M$.
- From the Taylor-expansion:

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

M-smooth functions

- f is *M-Lipschitz smooth*, S for $M > 0$ if
$$\nabla^2 f(x) \preceq M I_n.$$
- Equivalently: all eigenvalues of $\nabla^2 f(x)$ are $\leq M$.
- From the Taylor-expansion:
$$f(x) \leq f(y) + \nabla f(y)^\top (x - y) + \frac{M}{2}\|x - y\|^2.$$

Lemma

For $x \in S$,
$$f(x) - p^* \geq \frac{1}{2M}\|\nabla f(x)\|^2.$$

Condition number

- Let f be both *m-strongly convex* and *M-smooth* on S:

$$mI_n \preceq \nabla^2 f(x) \preceq MI_n.$$

- All eigenvalues of $\nabla^2 f(x)$ are in the interval $[m, M]$.
- Condition number

$$\kappa := \frac{M}{m}$$

- Important quantity to bound the convergence of gradient descent.

Convergence analysis of gradient descent

Theorem

Let f be m-strongly convex and M-smooth on S; let $\kappa = M/m$. Then, in the kth iteration of gradient descent with exact line search, we have

$$f(x^{(k)}) - p^* \leq \left(1 - \frac{1}{\kappa}\right)^k \cdot \left(f(x^{(0)}) - p^*\right).$$

Therefore, we find a solution $f(x^{(k)}) - p^* \leq \varepsilon$ within

$$k \leq \frac{\log \frac{f(x^{(0)}) - p^*}{\varepsilon}}{\log \frac{\kappa}{\kappa - 1}}$$

iterations. If κ is large, this is approximately

$$\kappa \log \frac{f(x^{(0)}) - p^*}{\varepsilon}.$$

Analysis for backtracking line search

Lemma

If f is M-smooth, then

$$f(x - t\nabla f(x)) \leq f(x) - \alpha t\|\nabla f(x)\|^2.$$

holds for every $\alpha \leq \frac{1}{2}$ and $0 \leq t \leq \frac{1}{M}$.

Analysis for backtracking line search

Lemma

If f is M-smooth, then

$$f(x - t\nabla f(x)) \leq f(x) - \alpha t \|\nabla f(x)\|^2.$$

holds for every $\alpha \leq \frac{1}{2}$ and $0 \leq t \leq \frac{1}{M}$.

Theorem

Let f be m-strongly convex and M-smooth on S. Then, in the kth iteration of gradient descent with backtracking line search, we have

$$f(x^{(k)}) - p^* \leq \left(1 - \frac{1}{\kappa'}\right)^k \cdot \left(f(x^{(0)}) - p^*\right)$$

for

$$\kappa' = \max\left\{\frac{1}{\alpha m}, \frac{M}{\alpha\beta m}\right\}.$$

Analysis for M-smooth functions

- Assume the function is not strongly convex, or we do not have a good bound $m > 0$.
- *Constant step size μ*, for $\mu \leq 1/M$.

$$f(x^{(k+1)}) \leq f(x^{(k)}) - \frac{\mu}{2}\|\nabla f(x^{(k)})\|^2.$$

Theorem

If f is M-smooth, then gradient descent with constant step size $\mu \leq 1/M$ finds a solution $x^{(k)}$ with $f(x_k) - p^* \leq \varepsilon$ for

$$k \leq \frac{C}{\mu\varepsilon}\|x^{(0)} - x^*\|^2$$

for some constant $C > 0$.

For $\mu = 1/M$,

$$k \leq \frac{CM}{\varepsilon}\|x^{(0)} - x^*\|^2.$$

Constrained optimisation

Frank-Wolfe algorithm = Conditional gradient method

- Convex set $X \subseteq \textbf{dom} f$

$$\begin{aligned} &\min f(x) \\ &\text{s. t. } x \in X. \end{aligned}$$

- Descent method using feasible points:

$$x^{(k+1)} = x^{(k)} + t_k \Delta x^{(k)}$$

such that $x^{(k+1)} \in X$ and $f(x^{(k+1)}) < f(x^{(k)})$.

- *Decreasing direction* $\Delta x^{(k)}$:

$$\nabla f(x^{(k)})^\top \Delta x^{(k)} < 0.$$

Frank-Wolfe algorithm = Conditional gradient method

- *Optimality criterion*

$$\nabla f(x^*)^\top (x - x^*) \geq 0 \text{ for all } x \in X.$$

- *Direction finding subproblem*:

$$\begin{aligned} &\min \nabla f(x^{(k)})^\top y \\ &\text{s. t. } y \in X. \end{aligned}$$

Frank-Wolfe algorithm = Conditional gradient method

- *Optimality criterion*

$$\nabla f(x^*)^\top (x - x^*) \geq 0 \text{ for all } x \in X.$$

- *Direction finding subproblem*:

$$\begin{aligned} \min\ & \nabla f(x^{(k)})^\top y \\ \text{s. t. } & y \in X. \end{aligned}$$

- If the optimum is $x^{(k)}$, we terminate.
- Otherwise, for optimum $s \in X$, we have

$$\nabla f(x^{(k)})^\top (s - x^{(k)}) < 0.$$

Frank-Wolfe algorithm = Conditional gradient method

- Decreasing direction $\Delta x^{(k)} = s - x^{(k)}$:

$$\nabla f(x^{(k)})^\top (s - x^{(k)}) < 0.$$

- Update

$$x^{(k+1)} = x^{(k)} + t_k(s - x^{(k)}) = (1 - t_k)x^{(k)} + t_k s$$

- $x^{(k+1)} \in X$ by convexity for every $0 \leq t_k \leq 1$.

Frank-Wolfe algorithm = Conditional gradient method

- Decreasing direction $\Delta x^{(k)} = s - x^{(k)}$:

$$\nabla f(x^{(k)})^\top (s - x^{(k)}) < 0.$$

- Update

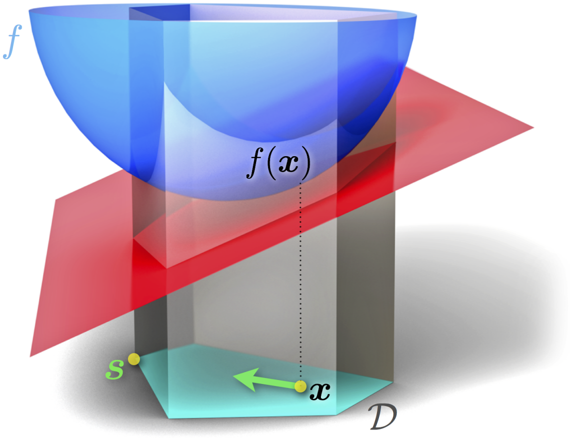
$$x^{(k+1)} = x^{(k)} + t_k(s - x^{(k)}) = (1 - t_k)x^{(k)} + t_k s$$

- $x^{(k+1)} \in X$ by convexity for every $0 \leq t_k \leq 1$.
- *Exact line search:*

$$t_k := \text{argmin}_{0 \leq t \leq 1} f((1 - t)x^{(k)} + ts).$$

- *Common choice:* $t_k = 2/(k + 1)$.

Frank-Wolfe algorithm = Conditional gradient method



Frank-Wolfe algorithm = Conditional gradient method

Lower bound

$$p^* = f(x^*) \geq f(x^{(k)}) + \nabla f(x^{(k)})^\top (s - x).$$

Frank-Wolfe algorithm = Conditional gradient method

Lower bound

$$p^* = f(x^*) \geq f(x^{(k)}) + \nabla f(x^{(k)})^\top (s - x).$$

Theorem

If f is M-smooth, then the Frank-Wolfe method with exact line search or with step size $t_k = 2/(k+1)$ finds a solution $x^{(k)}$ with $f(x_k) - p^* \leq \varepsilon$ for

$$k \leq \frac{CM}{\varepsilon} \|x^{(0)} - x^*\|^2$$

for some constant $C > 0$.