Gradient Methods

Goran Banjac

Large-Scale Convex Optimization
ETH Zurich

March 31, 2020

Lipschitz smoothness

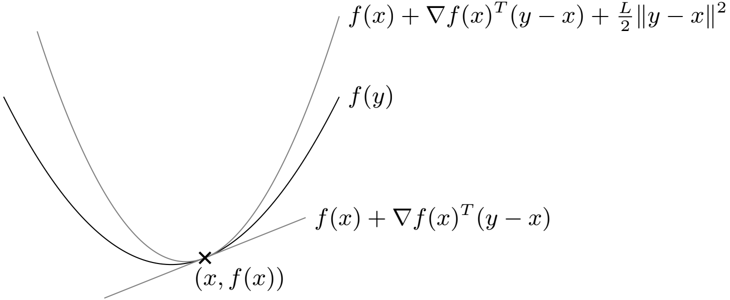
- a function $f\colon \mathbb{R}^n \mapsto \mathbb{R}$ is called M-*Lipschitz continuous* if for all $x, y \in \mathbb{R}^n$:
$$|f(x) - f(y)| \leq M\|x - y\|_2$$
- the slope of a Lipschitz continuous function is bounded
- f is called L-*Lipschitz smooth* (or just L-*smooth*) if it is differentiable and its gradient is L-Lipschitz continuous, *i.e.*, for all $x, y \in \mathbb{R}^n$:
$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2$$

Lipschitz smoothness of convex functions

- a convex function f is L-smooth if and only if

$$f(y) \leq f(x) + \nabla f(x)^T (y - x) + \frac{L}{2} \|y - x\|^2$$

for all $x, y \in \mathbb{R}^n$



- there exists a quadratic upper bound to f at every x

Majorization minimization

- *majorization minimization method* for solving

$$\text{minimize} \quad f(x)$$

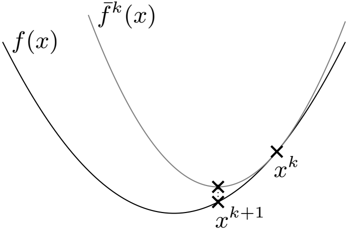
  - let the current iterate be x^k
  - find at x^k a majorizing function $\bar{f}^k$ such that

$$\bar{f}^k \geq f \quad \text{and} \quad \bar{f}^k(x^k) = f(x^k)$$

  - minimize $\bar{f}^k$ (easier than minimizing f) to get next iterate

$$x^{k+1} \in \underset{x}{\operatorname{argmin}}\, \bar{f}^k(x)$$

  - the majorizer should ensure $x^{k+1} = x^k$ if and only if x minimizes f



Gradient method

- let $f\colon \mathbb{R}^n \mapsto \mathbb{R}$ be a convex L-smooth function
- minimizing a quadratic majorizer if $\gamma_k \in [\varepsilon, \frac{1}{L}]$, $\varepsilon > 0$:

$$\begin{aligned} x^{k+1} &= \operatorname*{argmin}_y \left\{ f(x^k) + \nabla f(x^k)^T(y - x^k) + \tfrac{1}{2\gamma_k}\|y - x^k\|_2^2 \right\} \\ &= \operatorname*{argmin}_y \tfrac{1}{2\gamma_k}\|y - x^k + \gamma_k \nabla f(x^k)\|_2^2 \\ &= x^k - \gamma_k \nabla f(x^k) \end{aligned}$$

- gives *gradient method* with step size $\gamma_k \leq \frac{1}{L}$
- we will show later that larger steps are possible

Gradient method – fixed-point characterization

- denote $T_{\mathrm{GM}}^{\gamma} \coloneqq \mathrm{Id} - \gamma \nabla f$, where $\mathrm{Id} \colon x \mapsto x$ is the identity operator
- gradient method can be represented as iteration of operator T_{GM}^{γ}:

$$x^{k+1} = x^k - \gamma \nabla f(x^k) = T_{\mathrm{GM}}^{\gamma} x^k$$

- minimizer of f can be characterized as a fixed-point of T_{GM}^{γ}:

$$\operatorname{Fix} T_{\mathrm{GM}}^{\gamma} \coloneqq \{x \mid T_{\mathrm{GM}}^{\gamma} x = x\} = \{x \mid \nabla f(x) = 0\}$$

- does the fixed-point iteration converge to a fixed-point?

Function value decrease

- assume that $p^\star := \inf_x f(x) > -\infty$
- since f is L-smooth and $x^{k+1} = x^k - \gamma_k \nabla f(x^k)$, we have

$$
\begin{aligned}
f(x^{k+1}) &\le f(x^k) + \nabla f(x^k)^T (x^{k+1} - x^k) + \tfrac{L}{2}\|x^{k+1} - x^k\|_2^2 \\
&= f(x^k) - \tfrac{1}{\gamma_k}\|x^{k+1} - x^k\|_2^2 + \tfrac{L}{2}\|x^{k+1} - x^k\|_2^2 \\
&= f(x^k) - (\tfrac{1}{\gamma_k} - \tfrac{L}{2})\|x^{k+1} - x^k\|_2^2
\end{aligned}
$$

- the requirement on $\gamma_k \in [\varepsilon, \frac{2}{L} - \varepsilon]$, so that $\delta := \frac{1}{\gamma_k} - \frac{L}{2} > 0$
- function value will decrease as long as $x^{k+1} \neq x^k$

Convergence of fixed-point residual

- rearrange inequality from previous slide:

$$\delta \|x^{k+1} - x^k\|_2^2 \leq f(x^k) - f(x^{k+1})$$

- telescope summation gives for all $n \in \mathbb{N}$:

$$\begin{aligned} \delta \sum_{k=1}^{n} \|x^{k+1} - x^k\|_2^2 &\leq \sum_{k=1}^{n} \big(f(x^k) - f(x^{k+1})\big) \\ &= f(x^1) - f(x^{n+1}) \\ &\leq f(x^1) - p^\star < +\infty \end{aligned}$$

- since $\delta > 0$, this implies:

$$\|\nabla f(x^k)\|_2 = \tfrac{1}{\gamma_k} \|x^{k+1} - x^k\|_2 \to 0$$

- optimality condition is satisfied in the limit

Convergence rate

- convergence rate of gradient method is $\mathcal{O}(1/k)$
- if $\gamma = \frac{1}{L}$, then for all $k \geq 1$ and every solution $x^\star$:

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

- accelerated schemes exist that achieve optimal rate $\mathcal{O}(1/k^2)$
  - also known as *Nesterov's fast gradient method*
  - it adds a very specific varying momentum term to iterates
- if a function is in addition σ-strongly convex, then the convergence rate is linear (geometric) and depends on the condition number L/σ

Projection operator

- the *(Euclidean) projection* of $x \in \mathbb{R}^n$ on a nonempty closed convex set $\mathcal{C} \subseteq \mathbb{R}^n$ is defined as

$$\Pi_{\mathcal{C}}(x) \coloneqq \operatorname*{argmin}_{y \in \mathcal{C}} \|y - x\|_2 = \operatorname*{argmin}_{y} \left\{ \mathcal{I}_{\mathcal{C}}(y) + \tfrac{1}{2}\|y - x\|_2^2 \right\}$$

- using Fermat's rule and the identity $N_{\mathcal{C}} = \partial \mathcal{I}_{\mathcal{C}}$, we obtain

$$p = \Pi_{\mathcal{C}}(x) \quad \Longleftrightarrow \quad x - p \in N_{\mathcal{C}}(p)$$

Projected gradient method

- consider the constrained minimization problem

$$\text{minimize} \quad f(x) + \mathcal{I}_{\mathcal{C}}(x)$$

- let f be L-smooth and $\mathcal{C}$ a nonempty closed convex set
- *projected gradient method (PGM)* is given by

$$x^{k+1} = \Pi_{\mathcal{C}}\left(x^k - \gamma \nabla f(x^k)\right) =: T_{\mathrm{PGM}}^{\gamma} x^k$$

- a fixed-point of $T_{\mathrm{PGM}}^{\gamma}$ is a minimizer of the problem:

$$\operatorname{Fix} T_{\mathrm{PGM}}^{\gamma} = \{x \mid x = \Pi_{\mathcal{C}}\left(x - \gamma \nabla f(x)\right)\} = \{x \mid -\nabla f(x) \in N_{\mathcal{C}}(x)\}$$

- it is easy to show that the fixed-point residual of PGM converges to zero if $p^\star > -\infty$

Subgradient method

- assume f is closed and convex
- optimality condition:

$$x \in \underset{x}{\operatorname{argmin}}\, f(x) \quad \Longleftrightarrow \quad 0 \in \partial f(x) \quad \Longleftrightarrow \quad x \in x - \gamma \partial f(x)$$

- algorithm

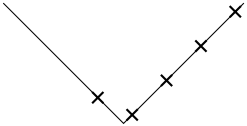
$$x^{k+1} \in x^k - \gamma \partial f(x^k)$$

- if we find a fixed-point, we solve the problem
- does the algorithm converge to a fixed-point?

Subgradient method – example

- consider minimizing the function $f(x) = |x|$
- let $\gamma = c$
- iteration if $x^k \neq nc$ where n is an integer:

$$x^{k+1} = x^k - c\,\mathrm{sign}(x)$$



- jumps back and forth over optimal point
- fixed step-size does not work

Convergence

- let $u^k \in \partial f(x^k)$ and $\|u\|_2 \leq G$ for all $u \in \partial f(x)$ and all x
- assume that there exists a minimizer $x^\star \in \operatorname{argmin}_x f(x)$
- from the subgradient definition, we have

$$p^\star = f(x^\star) \geq f(x^k) + (u^k)^T(x^\star - x^k)$$

- then

$$\begin{aligned} \|x^{k+1} - x^\star\|_2^2 &= \|x^k - \gamma_k u^k - x^\star\|_2^2 \\ &= \|x^k - x^\star\|_2^2 - 2\gamma_k (u^k)^T(x^k - x^\star) + \gamma_k^2 \|u^k\|_2^2 \\ &\leq \|x^k - x^\star\|_2^2 - 2\gamma_k (f(x^k) - p^\star) + \gamma_k^2 G^2 \end{aligned}$$

- telescope summation gives for all $n \in \mathbb{N}$:

$$\|x^{n+1} - x^\star\|_2^2 \leq \|x^0 - x^\star\|_2^2 - 2\sum_{k=0}^{n} \gamma_k (f(x^k) - p^\star) + G^2 \sum_{k=0}^{n} \gamma_k^2$$

Convergence

- let $f_{\text{best}}^n = \min_{k=1,\ldots,n} f(x^k)$; since $f(x^k) \geq p^\star$, we have

$$(f_{\text{best}}^n - p^\star) \sum_{k=0}^{n} \gamma_k = \sum_{k=0}^{n} \gamma_k (f_{\text{best}}^n - p^\star) \leq \sum_{k=0}^{n} \gamma_k (f(x^k) - p^\star)$$

- therefore

$$f_{\text{best}}^n - p^\star \leq \frac{\|x^0 - x^\star\|_2^2 + G^2 \sum_{k=0}^n \gamma_k^2}{2 \sum_{k=0}^n \gamma_k}$$

- if, for instance,

$$\sum_{k=0}^{\infty} \gamma_k = +\infty, \qquad \sum_{k=0}^{\infty} \gamma_k^2 < +\infty$$

  then numerator finite, but denominator $\to \infty$

- example: $\gamma_k = c/k$ for $c > 0$

Iterative methods for solving linear systems

- let $A \in \mathbb{S}^n_{++}$ be a symmetric positive definite matrix
- solution to the linear system:

$$Ax = b$$

  is equivalent to solution of the optimization problem:

$$\text{minimize} \quad f(x) \coloneqq \frac{1}{2}x^T Ax - b^T x$$

- $r = \nabla f(x) = Ax - b$ is the optimality residual
- hence, we can compute solution via gradient method
- convergence rate depends on $\mathrm{cond}(A) = \frac{\lambda_{\max}(A)}{\lambda_{\min}(A)}$

Conjugate direction method

- a set of nonzero vectors $\{p^0, \ldots, p^{n-1}\}$ is *conjugate* wrt A if

$$(p^i)^T A p^j = 0, \quad \forall i \neq j$$

- successive minimization of f along the conjugate directions:

$$\alpha^k = \operatorname*{argmin}_{\alpha} f(x^k + \alpha p^k) \quad \Longrightarrow \quad \alpha^k = -\frac{(r^k)^T p^k}{(p^k)^T A p^k}$$
$$x^{k+1} = x^k + \alpha^k p^k$$

- x^{k+1} minimizes f over the set $x^0 + S^k$, where S^k is given by

$$S^k = \operatorname{span}\{p^0, \ldots, p^k\}$$

- since the conjugate directions are linearly independent, the solution is computed in at most n iterations (in exact arithmetic)

Conjugate gradient method

- *conjugate gradient (CG) method* computes a set of conjugate directions efficiently
- p^k is computed as:
$$p^k = -r^k + \beta^k p^{k-1}$$
- initial direction is set to $p^0 = -r^0$
- β^k is computed from the conjugacy requirement $(p^k)^T A p^{k-1} = 0$:
$$\beta^k = \frac{(r^k)^T A p^{k-1}}{(p^{k-1})^T A p^{k-1}}$$
- a good approximate solution can be obtained after $d \ll n$ iterations
- preconditioning improves the convergence rate
- *warm-starting* can speed-up convergence considerably

References

- these lecture notes are based to a large extent on the following material:
  - Stanford EE364b class developed by Stephen Boyd
  - Lund course on Large-Scale Convex Optimization developed by Pontus Giselsson

- the original slides can be downloaded from

  `https://web.stanford.edu/class/ee364b/lectures.html`

  `https://archive.control.lth.se/ls-convex-2015/`