Optimization for Machine Learning

Lecture 8: Subgradient method; Accelerated gradient

6.881: MIT

Suvrit Sra
Massachusetts Institute of Technology

16 Mar, 2021

First-order methods

$$x \leftarrow x - \eta g(x)$$

$\nabla f(x)$

GD

$\mathbb{E}[g(x)] = \nabla f(x)$

SGD

$g(x) \in \partial f(x)$

Subgrad

Subgradient method

Unconstrained convex problem

$$\min_x \quad f(x)$$

Unconstrained convex problem

$$\min_{x} \quad f(x)$$

1. Start with some guess x^0; set $k = 0$

Unconstrained convex problem

$$\min_x \quad f(x)$$

1. Start with some guess x^0; set $k = 0$
2. If $0 \in \partial f(x^k)$, **stop**; output x^k

Unconstrained convex problem

$$\min_x \quad f(x)$$

1. Start with some guess x^0; set $k = 0$
2. If $0 \in \partial f(x^k)$, **stop**; output x^k
3. Otherwise, generate next guess x^{k+1}

Unconstrained convex problem

$$\min_x \quad f(x)$$

1. Start with some guess x^0; set $k = 0$
2. If $0 \in \partial f(x^k)$, **stop**; output x^k
3. Otherwise, generate next guess x^{k+1}
4. Repeat above procedure until $f(x^k) \leq f(x^*) + \varepsilon$

Subgradient method

$$x^{k+1} = x^k - \eta_k g^k$$

where $g^k \in \partial f(x^k)$ is **any** subgradient

Subgradient method

$$x^{k+1} = x^k - \eta_k g^k$$

where $g^k \in \partial f(x^k)$ is **any** subgradient

Stepsize $\eta_k > 0$ must be chosen

Subgradient method

$$x^{k+1} = x^k - \eta_k g^k$$

where $g^k \in \partial f(x^k)$ is **any** subgradient

Stepsize $\eta_k > 0$ must be chosen

- Method generates sequence $\{x^k\}_{k \geq 0}$
- Does this sequence converge to an optimal solution x^*?
- If yes, then how fast?
- What if we have constraints: $x \in \mathcal{C}$?

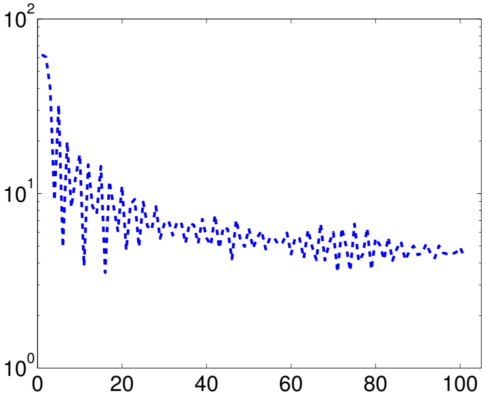
Example

$$\min \quad \tfrac{1}{2}\|Ax - b\|_2^2 + \lambda\|x\|_1$$
$$x^{k+1} = x^k - \eta_k(A^T(Ax^k - b) + \lambda\,\mathrm{sgn}(x^k))$$

Example

$$\min \quad \tfrac{1}{2}\|Ax - b\|_2^2 + \lambda\|x\|_1$$

$$x^{k+1} = x^k - \eta_k(A^T(Ax^k - b) + \lambda\,\mathrm{sgn}(x^k))$$



Example

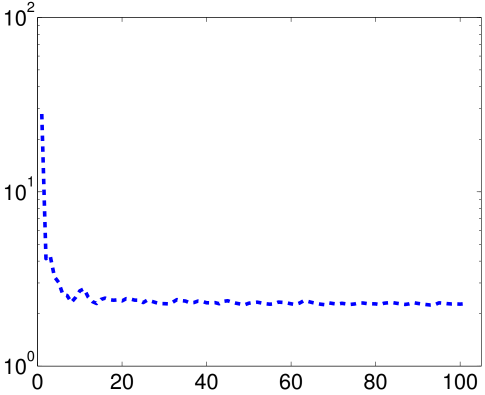
$$\min \quad \tfrac{1}{2}\|Ax - b\|_2^2 + \lambda\|x\|_1$$

$$x^{k+1} = x^k - \eta_k(A^T(Ax^k - b) + \lambda\,\mathrm{sgn}(x^k))$$

Example

$$\min \quad \tfrac{1}{2}\|Ax - b\|_2^2 + \lambda\|x\|_1$$

$$x^{k+1} = x^k - \eta_k(A^T(Ax^k - b) + \lambda\,\mathrm{sgn}(x^k))$$



(More careful implementation)

Exercise

Exercise: Experiment with deep neural network classifier where we want to learn *sparse* weights. In particular, experiment with the following loss function:

$$\min_x \; L(x) := \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, \mathcal{NN}(x, a_i)) + \lambda\|x\|_1.$$

Implement a stochastic subgradient update to minimize L.
(*Hint:* If we pretend that the loss part is differentiable, then we can invoke Clarke's rule: $\partial_\circ L = \nabla\text{loss} + \lambda\partial\text{reg}$)

Subgradient method – stepsizes

- **Constant** Set $\eta_k = \eta > 0$, for $k \geq 0$
- **Normalized** $\eta_k = \eta / \|g^k\|_2 \quad (\|x^{k+1} - x^k\|_2 = \eta)$
- **Square summable**

$$\sum_k \eta_k^2 < \infty, \qquad \sum_k \eta_k = \infty$$

- **Diminishing**

$$\lim_k \eta_k = 0, \qquad \sum_k \eta_k = \infty$$

- **Adaptive stepsizes** (not covered)

Subgradient method – stepsizes

- **Constant** Set $\eta_k = \eta > 0$, for $k \geq 0$
- **Normalized** $\eta_k = \eta/\|g^k\|_2 \quad (\|x^{k+1} - x^k\|_2 = \eta)$
- **Square summable**

$$\sum_k \eta_k^2 < \infty, \qquad \sum_k \eta_k = \infty$$

- **Diminishing**

$$\lim_k \eta_k = 0, \qquad \sum_k \eta_k = \infty$$

- **Adaptive stepsizes** (not covered)

Not a descent method!
Could use best f^k so far: $f^k_{\min} := \min_{0 \leq i \leq k} f^i$

Convergence

(sketch)

Convergence analysis

Assumptions

- Min is attained: $f^* := \inf_x f(x) > -\infty$, with $f(x^*) = f^*$

Convergence analysis

Assumptions

- Min is attained: $f^* := \inf_x f(x) > -\infty$, with $f(x^*) = f^*$
- Bounded subgradients: $\|g\|_2 \le G$ for all $g \in \partial f$

Convergence analysis

Assumptions

- Min is attained: $f^* := \inf_x f(x) > -\infty$, with $f(x^*) = f^*$
- Bounded subgradients: $\|g\|_2 \le G$ for all $g \in \partial f$
- Bounded domain: $\|x^0 - x^*\|_2 \le R$

Convergence analysis

Assumptions

- Min is attained: $f^* := \inf_x f(x) > -\infty$, with $f(x^*) = f^*$
- Bounded subgradients: $\|g\|_2 \leq G$ for all $g \in \partial f$
- Bounded domain: $\|x^0 - x^*\|_2 \leq R$

Convergence results for: $f^k_{\min} := \min_{0 \leq i \leq k} f^i$

Subgradient method – convergence

Lyapunov function: Distance to x^* (instead of $f - f^*$)

Subgradient method – convergence

Lyapunov function: **Distance to x^*** (instead of $f - f^*$)

$$\|x^{k+1} - x^*\|_2^2 \quad = \quad \|x^k - \eta_k g^k - x^*\|_2^2$$

Subgradient method – convergence

Lyapunov function: Distance to x^* (instead of $f - f^*$)

$$
\begin{aligned}
\|x^{k+1} - x^*\|_2^2 &= \|x^k - \eta_k g^k - x^*\|_2^2 \\
&= \|x^k - x^*\|_2^2 + \eta_k^2 \|g^k\|_2^2 - 2\langle \eta_k g^k,\, x^k - x^* \rangle
\end{aligned}
$$

Subgradient method – convergence

Lyapunov function: **Distance to x^*** (instead of $f - f^*$)

$$
\begin{aligned}
\|x^{k+1} - x^*\|_2^2 &= \|x^k - \eta_k g^k - x^*\|_2^2 \\
&= \|x^k - x^*\|_2^2 + \eta_k^2 \|g^k\|_2^2 - 2\langle \eta_k g^k,\, x^k - x^* \rangle \\
&\leq \|x^k - x^*\|_2^2 + \eta_k^2 \|g^k\|_2^2 - 2\eta_k (f(x^k) - f^*),
\end{aligned}
$$

since $f^* = f(x^*) \geq f(x^k) + \langle g^k,\, x^* - x^k \rangle$

Subgradient method – convergence

Lyapunov function: Distance to x^* (instead of $f - f^*$)

$$\begin{aligned} \|x^{k+1} - x^*\|_2^2 &= \|x^k - \eta_k g^k - x^*\|_2^2 \\ &= \|x^k - x^*\|_2^2 + \eta_k^2 \|g^k\|_2^2 - 2\langle \eta_k g^k, x^k - x^* \rangle \\ &\le \|x^k - x^*\|_2^2 + \eta_k^2 \|g^k\|_2^2 - 2\eta_k (f(x^k) - f^*), \end{aligned}$$

since $f^* = f(x^*) \ge f(x^k) + \langle g^k, x^* - x^k \rangle$

Apply same argument to $\|x^k - x^*\|_2^2$ recursively

Subgradient method – convergence

Lyapunov function: **Distance to x^*** (instead of $f - f^*$)

$$\begin{aligned} \|x^{k+1} - x^*\|_2^2 &= \|x^k - \eta_k g^k - x^*\|_2^2 \\ &= \|x^k - x^*\|_2^2 + \eta_k^2 \|g^k\|_2^2 - 2\langle \eta_k g^k, x^k - x^* \rangle \\ &\leq \|x^k - x^*\|_2^2 + \eta_k^2 \|g^k\|_2^2 - 2\eta_k (f(x^k) - f^*), \end{aligned}$$

since $f^* = f(x^*) \geq f(x^k) + \langle g^k, x^* - x^k \rangle$

Apply same argument to $\|x^k - x^*\|_2^2$ recursively

$$\|x^{k+1} - x^*\|_2^2 \leq \|x^0 - x^*\|_2^2 + \sum_{t=1}^{k} \eta_t^2 \|g^t\|_2^2 - 2\sum_{t=1}^{k} \eta_t (f^t - f^*).$$

Subgradient method – convergence

Lyapunov function: Distance to x^* (instead of $f - f^*$)

$$\begin{aligned}\|x^{k+1} - x^*\|_2^2 &= \|x^k - \eta_k g^k - x^*\|_2^2 \\ &= \|x^k - x^*\|_2^2 + \eta_k^2 \|g^k\|_2^2 - 2\langle \eta_k g^k, x^k - x^* \rangle \\ &\leq \|x^k - x^*\|_2^2 + \eta_k^2 \|g^k\|_2^2 - 2\eta_k (f(x^k) - f^*),\end{aligned}$$

since $f^* = f(x^*) \geq f(x^k) + \langle g^k, x^* - x^k \rangle$

Apply same argument to $\|x^k - x^*\|_2^2$ recursively

$$\|x^{k+1} - x^*\|_2^2 \leq \|x^0 - x^*\|_2^2 + \sum_{t=1}^{k} \eta_t^2 \|g^t\|_2^2 - 2 \sum_{t=1}^{k} \eta_t (f^t - f^*).$$

Now use our convenient assumptions!

Subgradient method – convergence

$$\|x^{k+1} - x^*\|_2^2 \le R^2 + G^2 \sum_{t=1}^{k} \eta_t^2 - 2\sum_{t=1}^{k} \eta_t (f^t - f^*).$$

- To get a bound on the last term, simply notice (for $t \le k$)

$$f^t \ge f^t_{\min} \ge f^k_{\min} \qquad \text{since} \quad f^t_{\min} := \min_{0 \le i \le t} f(x^i)$$

Subgradient method – convergence

$$\|x^{k+1} - x^*\|_2^2 \leq R^2 + G^2 \sum_{t=1}^{k} \eta_t^2 - 2\sum_{t=1}^{k} \eta_t (f^t - f^*).$$

- To get a bound on the last term, simply notice (for $t \leq k$)

$$f^t \geq f_{\min}^t \geq f_{\min}^k \qquad \text{since} \quad f_{\min}^t := \min_{0 \leq i \leq t} f(x^i)$$

- Plugging this in yields the bound

$$2\sum_{t=1}^{k} \eta_t (f^t - f^*) \geq 2(f_{\min}^k - f^*) \sum_{t=1}^{k} \eta_t.$$

Subgradient method – convergence

$$\|x^{k+1} - x^*\|_2^2 \le R^2 + G^2 \sum_{t=1}^{k} \eta_t^2 - 2 \sum_{t=1}^{k} \eta_t (f^t - f^*).$$

- To get a bound on the last term, simply notice (for $t \le k$)

$$f^t \ge f_{\min}^t \ge f_{\min}^k \qquad \text{since} \quad f_{\min}^t := \min_{0 \le i \le t} f(x^i)$$

- Plugging this in yields the bound

$$2 \sum_{t=1}^{k} \eta_t (f^t - f^*) \ge 2(f_{\min}^k - f^*) \sum_{t=1}^{k} \eta_t.$$

- So that we finally have

$$0 \le \|x^{k+1} - x^*\|_2 \le R^2 + G^2 \sum_{t=1}^{k} \eta_t^2 - 2(f_{\min}^k - f^*) \sum_{t=1}^{k} \eta_t$$

Subgradient method – convergence

$$\|x^{k+1} - x^*\|_2^2 \leq R^2 + G^2 \sum_{t=1}^{k} \eta_t^2 - 2\sum_{t=1}^{k} \eta_t (f^t - f^*).$$

- To get a bound on the last term, simply notice (for $t \leq k$)

$$f^t \geq f_{\min}^t \geq f_{\min}^k \qquad \text{since} \quad f_{\min}^t := \min_{0 \leq i \leq t} f(x^i)$$

- Plugging this in yields the bound

$$2\sum_{t=1}^{k} \eta_t (f^t - f^*) \geq 2(f_{\min}^k - f^*) \sum_{t=1}^{k} \eta_t.$$

- So that we finally have

$$0 \leq \|x^{k+1} - x^*\|_2 \leq R^2 + G^2 \sum_{t=1}^{k} \eta_t^2 - 2(f_{\min}^k - f^*) \sum_{t=1}^{k} \eta_t$$

$$\boxed{f_{\min}^k - f^* \leq \frac{R^2 + G^2 \sum_{t=1}^{k} \eta_t^2}{2\sum_{t=1}^{k} \eta_t}}$$

Subgradient method – convergence

$$f_{\min}^k - f^* \leq \frac{R^2 + G^2 \sum_{t=1}^k \eta_t^2}{2\sum_{t=1}^k \eta_t}$$

Exercise: Analyze $\lim_{k\to\infty} f_{\min}^k - f^*$ for the different choices of stepsize that we mentioned.

Subgradient method – convergence

$$f_{\min}^k - f^* \leq \frac{R^2 + G^2 \sum_{t=1}^k \eta_t^2}{2 \sum_{t=1}^k \eta_t}$$

Exercise: Analyze $\lim_{k \to \infty} f_{\min}^k - f^*$ for the different choices of stepsize that we mentioned.

Constant step: $\eta_k = \eta$; We obtain

$$f_{\min}^k - f^* \leq \frac{R^2 + G^2 k \eta^2}{2k\eta}$$

Subgradient method – convergence

$$f_{\min}^k - f^* \le \frac{R^2 + G^2 \sum_{t=1}^k \eta_t^2}{2 \sum_{t=1}^k \eta_t}$$

Exercise: Analyze $\lim_{k\to\infty} f_{\min}^k - f^*$ for the different choices of stepsize that we mentioned.

Constant step: $\eta_k = \eta$; We obtain

$$f_{\min}^k - f^* \le \frac{R^2 + G^2 k \eta^2}{2k\eta} \to \frac{G^2\eta}{2} \quad \text{as } k \to \infty.$$

Subgradient method – convergence

$$f_{\min}^k - f^* \leq \frac{R^2 + G^2 \sum_{t=1}^k \eta_t^2}{2 \sum_{t=1}^k \eta_t}$$

Exercise: Analyze $\lim_{k\to\infty} f_{\min}^k - f^*$ for the different choices of stepsize that we mentioned.

Constant step: $\eta_k = \eta$; We obtain

$$f_{\min}^k - f^* \leq \frac{R^2 + G^2 k \eta^2}{2k\eta} \to \frac{G^2 \eta}{2} \quad \text{as } k \to \infty.$$

Square summable, not summable: $\sum_k \eta_k^2 < \infty$, $\sum_k \eta_k = \infty$

Subgradient method – convergence

$$f_{\min}^k - f^* \leq \frac{R^2 + G^2 \sum_{t=1}^k \eta_t^2}{2\sum_{t=1}^k \eta_t}$$

Exercise: Analyze $\lim_{k\to\infty} f_{\min}^k - f^*$ for the different choices of stepsize that we mentioned.

Constant step: $\eta_k = \eta$; We obtain

$$f_{\min}^k - f^* \leq \frac{R^2 + G^2 k \eta^2}{2k\eta} \to \frac{G^2\eta}{2} \quad \text{as } k \to \infty.$$

Square summable, not summable: $\sum_k \eta_k^2 < \infty$, $\sum_k \eta_k = \infty$

As $k \to \infty$, numerator $< \infty$ but denominator $\to \infty$; so $f_{\min}^k \to f^*$

Subgradient method – convergence

$$f_{\min}^k - f^* \leq \frac{R^2 + G^2 \sum_{t=1}^k \eta_t^2}{2 \sum_{t=1}^k \eta_t}$$

Exercise: Analyze $\lim_{k\to\infty} f_{\min}^k - f^*$ for the different choices of stepsize that we mentioned.

Constant step: $\eta_k = \eta$; We obtain

$$f_{\min}^k - f^* \leq \frac{R^2 + G^2 k \eta^2}{2k\eta} \to \frac{G^2\eta}{2} \quad \text{as } k \to \infty.$$

Square summable, not summable: $\sum_k \eta_k^2 < \infty$, $\sum_k \eta_k = \infty$

As $k \to \infty$, numerator $< \infty$ but denominator $\to \infty$; so $f_{\min}^k \to f^*$

In practice, fair bit of stepsize tuning needed, e.g. $\eta_t = a/(b+t)$

Subgradient method – convergence

- Suppose we want $f_{\min}^k - f^* \le \varepsilon$, how big should k be?

Subgradient method – convergence

- Suppose we want $f_{\min}^k - f^* \leq \varepsilon$, how big should k be?
- Optimize the bound for η_t: want

$$f_{\min}^k - f^* \leq \qquad\qquad\qquad\qquad \varepsilon$$

Subgradient method – convergence

- Suppose we want $f_{\min}^k - f^* \leq \varepsilon$, how big should k be?
- Optimize the bound for η_t: want

$$f_{\min}^k - f^* \leq \frac{R^2 + G^2 \sum_{t=1}^k \eta_t^2}{2 \sum_{t=1}^k \eta_t} \quad \leq \quad \varepsilon$$

Subgradient method – convergence

- Suppose we want $f_{\min}^k - f^* \leq \varepsilon$, how big should k be?
- Optimize the bound for η_t: want

$$f_{\min}^k - f^* \leq \frac{R^2 + G^2 \sum_{t=1}^k \eta_t^2}{2 \sum_{t=1}^k \eta_t} \quad \leq \quad \varepsilon$$

- For fixed k: best possible stepsize is constant η

$$\frac{R^2 + G^2 k \eta^2}{2k\eta} \leq \epsilon \quad \Rightarrow \quad \eta = \frac{R}{G\sqrt{k}}$$

Subgradient method – convergence

- Suppose we want $f_{\min}^k - f^* \leq \varepsilon$, how big should k be?
- Optimize the bound for η_t: want

$$f_{\min}^k - f^* \leq \frac{R^2 + G^2 \sum_{t=1}^k \eta_t^2}{2 \sum_{t=1}^k \eta_t} \leq \varepsilon$$

- For fixed k: best possible stepsize is constant η

$$\frac{R^2 + G^2 k \eta^2}{2k\eta} \leq \epsilon \quad \Rightarrow \quad \eta = \frac{R}{G\sqrt{k}}$$

- Then, after k steps $f_{\min}^k - f^* \leq RG/\sqrt{k}$.
- For accuracy ϵ, we need at least $(RG/\epsilon)^2 = O(1/\epsilon^2)$ steps

Subgradient method – convergence

- Suppose we want $f_{\min}^k - f^* \leq \varepsilon$, how big should k be?
- Optimize the bound for η_t: want

$$f_{\min}^k - f^* \leq \frac{R^2 + G^2 \sum_{t=1}^k \eta_t^2}{2\sum_{t=1}^k \eta_t} \leq \varepsilon$$

- For fixed k: best possible stepsize is constant η

$$\frac{R^2 + G^2 k \eta^2}{2k\eta} \leq \epsilon \quad \Rightarrow \quad \eta = \frac{R}{G\sqrt{k}}$$

- Then, after k steps $f_{\min}^k - f^* \leq RG/\sqrt{k}$.
- For accuracy ϵ, we need at least $(RG/\epsilon)^2 = O(1/\epsilon^2)$ steps
- (quite slow **but already hits the lower bound!**)

Exercise: Support vector machines

- Let $\mathcal{D} := \{(x_i, y_i) \mid x_i \in \mathbb{R}^n, y_i \in \{\pm 1\}\}$
- We wish to find $w \in \mathbb{R}^n$ and $b \in \mathbb{R}$ such that

$$\min_{w,b} \quad \tfrac{1}{2}\|w\|_2^2 + C\sum\nolimits_{i=1}^{m} \max[0, 1 - y_i(w^T x_i + b)]$$

- Derive and implement a subgradient method
- Plot evolution of objective function
- Experiment with different values of $C > 0$
- Plot and keep track of $f_{\min}^k := \min_{0 \le t \le k} f(x^t)$

Exercise: Geometric median

- Let $a \in \mathbb{R}^n$ be a given vector.
- Let $f(x) = \sum_i |x - a_i|$, i.e., $f : \mathbb{R} \to \mathbb{R}_+$
- Implement different subgradient methods to minimize f
- Also keep track of $f_{\text{best}}^k := \min_{0 \le i < k} f(x_i)$

Exercise: Implement the above. Plot the $f(x_k)$ values; also try to guess what optimum is being found.

Optimization with simple constraints

$$\min \quad f(x) \qquad \text{s.t.} \quad x \in \mathcal{C}$$

Optimization with simple constraints

$$\min \quad f(x) \qquad \text{s.t.} \quad x \in \mathcal{C}$$

- Previously:

$$x^{t+1} = x^t - \eta_t g^t$$

- This could be infeasible!

Optimization with simple constraints

$$\min \quad f(x) \qquad \text{s.t.} \quad x \in \mathcal{C}$$

- Previously:

$$x^{t+1} = x^t - \eta_t g^t$$

- This could be infeasible!
- **Use projection**

Projected subgradient method

$$x^{k+1} = P_{\mathcal{C}}(x^k - \eta_k g^k)$$

where $g^k \in \partial f(x^k)$ is any subgradient

Projected subgradient method

$$x^{k+1} = P_{\mathcal{C}}(x^k - \eta_k g^k)$$

where $g^k \in \partial f(x^k)$ is any subgradient

- **Projection** closest feasible point

$$P_{\mathcal{C}}(x) = \operatorname*{arg\,min}_{y \in \mathcal{C}} \|x - y\|^2$$

(Assume $\mathcal{C}$ is closed and convex, then projection is unique)

Projected subgradient method

$$x^{k+1} = P_{\mathcal{C}}(x^k - \eta_k g^k)$$

where $g^k \in \partial f(x^k)$ is any subgradient

- **Projection** closest feasible point

$$P_{\mathcal{C}}(x) = \operatorname*{arg\,min}_{y \in \mathcal{C}} \|x - y\|^2$$

(Assume $\mathcal{C}$ is closed and convex, then projection is unique)

- Great as long as projection is "easy"
- Same questions as before:
  - Does it converge? For which stepsizes? How fast?

Key idea: Projection Theorem

Let $\mathcal{C}$ be nonempty, closed and convex.

- **Recall:** Optimality conditions: $y^* = P_{\mathcal{C}}(z)$ iff

$$\langle z - y^*, y - y^* \rangle \leq 0 \text{ for all } y \in \mathcal{C}$$

Verify: Projection is nonexpansive:

$$\|P_{\mathcal{C}}(x) - P_{\mathcal{C}}(z)\| \;\leq\; \|x - z\|^2 \quad \text{for all } x, z \in \mathbb{R}^n.$$

Convergence analysis

Assumptions

- Min is attained: $f^* := \inf_x f(x) > -\infty$, with $f(x^*) = f^*$
- Bounded subgradients: $\|g\|_2 \le G$ for all $g \in \partial f$
- Bounded domain: $\|x^0 - x^*\|_2 \le R$

Analysis

- Let $z^{t+1} = P_{\mathcal{C}}(x^t - \eta_t g^t)$.
- Then $x^{t+1} = P_{\mathcal{C}}(z^{t+1})$.
- Recall analysis of unconstrained method:

$$
\begin{aligned}
\|z^{t+1} - x^*\|_2^2 &= \|x^t - \eta_t g^t - x^*\|_2^2 \\
&\le \|x^t - x^*\|_2^2 + \eta_t^2 \|g^t\|_2^2 - 2\eta_t (f(x^t) - f^*) \\
&\dots
\end{aligned}
$$

- Need to relate to $\|x^{t+1} - x^*\|_2^2$, the rest is as before

Convergence analysis: Key idea

- Using nonexpansiveness of projection:

$$
\begin{aligned}
&\|x^t - \eta_t g^t - x^*\|_2^2 \\
\leq\ &\|x^t - x^*\|_2^2 + \eta_t^2 \|g^t\|_2^2 - 2\eta_t (f(x^t) - f^*) \\
&\ldots
\end{aligned}
$$

Convergence analysis: Key idea

- Using nonexpansiveness of projection:

$$\begin{aligned}
\|x^{t+1} - x^*\|_2^2 &= \|P_C(x^t - \eta_t g^t) - P_C(x^*)\|_2^2 \\
&\le \|x^t - \eta_t g^t - x^*\|_2^2 \\
&\le \|x^t - x^*\|_2^2 + \eta_t^2 \|g^t\|_2^2 - 2\eta_t (f(x^t) - f^*) \\
&\ldots
\end{aligned}$$

Same convergence results as in unconstrained case:

- within neighborhood of optimal for constant step size
- converges for diminishing non-summable

Examples of simple projections

- **Nonnegativity** $x \geq 0$, $P_{\mathcal{C}}(z) = [z]_+$
- **ℓ_∞-ball** $\|x\|_\infty \leq 1$
  Projection: $\min \|x - z\|^2$ s.t. $x \leq 1$ and $x \geq -1$
  $P_{\|x\|_\infty \leq 1}(z) = y$ where $y_i = \operatorname{sgn}(z_i) \min\{|z_i|, 1\}$
- **Linear equality constraints** $Ax = b$ ($A \in \mathbb{R}^{n \times m}$ has rank n)

$$\begin{aligned} P_{\mathcal{C}}(x) &= z - A^\top (AA^\top)^{-1}(Az - b) \\ &= (I - A^\top (A^\top A)^{-1} A)z + A^\top (AA^\top)^{-1} b \end{aligned}$$

- **Simplex:** $x^\top 1 = 1$ and $x \geq 0$
  doable in $O(n)$ time; similarly ℓ_1-norm ball

Some remarks

- Why care?
  - simple
  - low-memory
  - stochastic version possible

Some remarks

- Why care?
  - simple
  - low-memory
  - stochastic version possible

Another perspective

$$x^{k+1} = \min_{x \in \mathcal{C}} \langle x, g^k \rangle + \frac{1}{2\eta_k} \|x - x_k\|^2$$

Mirror Descent version

$$x^{k+1} = \min_{x \in \mathcal{C}} \langle x, g^k \rangle + \frac{1}{\eta_k} D_\varphi(x, x_k)$$

Accelerated gradient

Gradient methods – upper bounds

Theorem. (Upper bound I). Let $f \in C^1_L$. Then,

$$\min_k \|\nabla f(x^k)\| \leq \varepsilon \text{ in } O(1/\varepsilon^2) \text{ iterations.}$$

Gradient methods – upper bounds

Theorem. (Upper bound I). Let $f \in C^1_L$. Then,

$$\min_k \|\nabla f(x^k)\| \leq \varepsilon \text{ in } O(1/\varepsilon^2) \text{ iterations.}$$

Theorem. (Upper bound II). Let $f \in S^1_{L,\mu}$. Then,

$$f(x^k) - f(x^*) \leq \frac{L}{2}\left(\frac{\kappa-1}{\kappa+1}\right)^{2k} \|x^0 - x^*\|_2^2$$

Gradient methods – upper bounds

Theorem. (Upper bound I). Let $f \in C^1_L$. Then,

$$\min_k \|\nabla f(x^k)\| \leq \varepsilon \text{ in } O(1/\varepsilon^2) \text{ iterations.}$$

Theorem. (Upper bound II). Let $f \in S^1_{L,\mu}$. Then,

$$f(x^k) - f(x^*) \leq \frac{L}{2}\left(\frac{\kappa - 1}{\kappa + 1}\right)^{2k} \|x^0 - x^*\|_2^2$$

Theorem. (Upper bound III). Let $f \in C^1_L$ be convex. Then,

$$f(x^k) - f(x^*) \leq \frac{2L(f(x^0) - f(x^*))\|x^0 - x^*\|_2^2}{k + 4}.$$

Gradient methods – lower bounds

Theorem. (Carmon-Duchi-Hinder-Sidford 2017). There's an $f \in C^1_L$, such that $\|\nabla f(x)\| \le \varepsilon$ requires $\Omega(\varepsilon^{-2})$ gradient evaluations.

Gradient methods – lower bounds

Theorem. (Carmon-Duchi-Hinder-Sidford 2017). There's an $f \in C_L^1$, such that $\|\nabla f(x)\| \leq \varepsilon$ requires $\Omega(\varepsilon^{-2})$ gradient evaluations.

Theorem. (Nesterov). There exists $f \in S_{L,\mu}^{\infty}$ ($\mu > 0$, $\kappa > 1$) s.t.

$$f(x^k) - f(x^*) \quad \geq \quad \frac{\mu}{2}\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{2k}\|x^0 - x^*\|_2^2,$$

Gradient methods – lower bounds

Theorem. (Carmon-Duchi-Hinder-Sidford 2017). There's an $f \in C_L^1$, such that $\|\nabla f(x)\| \leq \varepsilon$ requires $\Omega(\varepsilon^{-2})$ gradient evaluations.

Theorem. (Nesterov). There exists $f \in S_{L,\mu}^{\infty}$ ($\mu > 0$, $\kappa > 1$) s.t.

$$f(x^k) - f(x^*) \quad \geq \quad \frac{\mu}{2}\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{2k} \|x^0 - x^*\|_2^2,$$

Theorem. (Nesterov). For any $x^0 \in \mathbb{R}^n$, and $1 \leq k \leq \frac{1}{2}(n-1)$, there is a convex $f \in C_L^1$, s.t.

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

$$\|x^k - x^0\|^2 \quad \geq \quad \tfrac{1}{8}\|x^0 - x^*\|^2.$$

Accelerated gradient methods

Upper bounds: (i) $O(1/k)$; and (ii) linear rate involving κ

Lower bounds: (i) $O(1/k^2)$; and (ii) linear rate involving $\sqrt{\kappa}$

Challenge: Close this gap!

Accelerated gradient methods

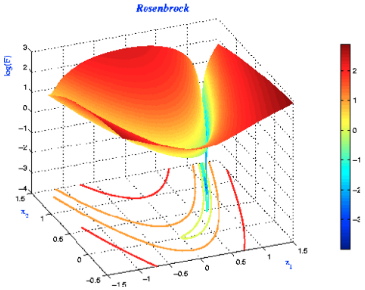
Upper bounds: (i) $O(1/k)$; and (ii) linear rate involving κ

Lower bounds: (i) $O(1/k^2)$; and (ii) linear rate involving $\sqrt{\kappa}$

Challenge: Close this gap!

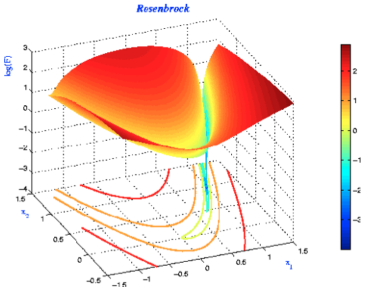
Nesterov (1983) closed the gap.

Background: ravine method



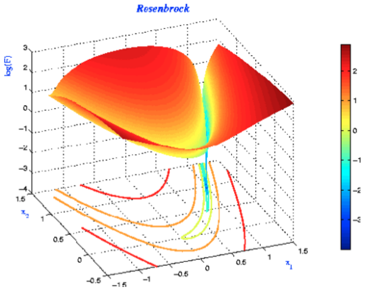
- Long, narrow **ravines** slow down GD

Background: ravine method



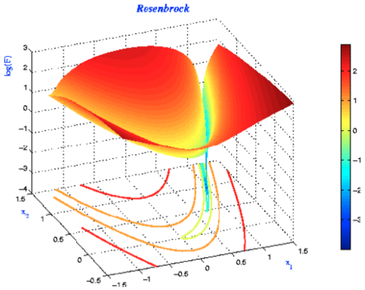
- Long, narrow **ravines** slow down GD
- Gel'fand-Tsetlin (1961): *Ravine method*

Background: ravine method



- Long, narrow **ravines** slow down GD
- Gel'fand-Tsetlin (1961): *Ravine method*
- Intuition: descent to bottom of ravine not hard, but moving along narrow ravine harder. Thus, *mix two types of steps*: gradient step and a "ravine step"

Background: ravine method



- Long, narrow **ravines** slow down GD
- Gel'fand-Tsetlin (1961): *Ravine method*
- Intuition: descent to bottom of ravine not hard, but moving along narrow ravine harder. Thus, *mix two types of steps*: gradient step and a "ravine step"

Simplest form of ravine method

$$x^{k+1} = y^k - \alpha \nabla f(y^k), \quad y^{k+1} = x^{k+1} + \beta(x^{k+1} - x^k)$$

Background: Heavy-ball method

Polyak's Momentum Method (1964)

$$x^{k+1} = x^k - \eta_k \nabla f(x^k) + \beta_k (x^k - x^{k-1})$$

Theorem. Let $f = \frac{1}{2}x^T Ax + b^T x \in S^1_{L,\mu}$. Then, choose

$$\eta_k = 4/(\sqrt{L} + \sqrt{\mu}), \quad \beta_k = q^2, q = \frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1}$$

the heavy-ball method satisfies $\|x^k - x^*\| = O(q^k)$.

Background: Heavy-ball method

Polyak's Momentum Method (1964)

$$x^{k+1} = x^k - \eta_k \nabla f(x^k) + \beta_k(x^k - x^{k-1})$$

Theorem. Let $f = \frac{1}{2}x^T A x + b^T x \in S^1_{L,\mu}$. Then, choose

$$\eta_k = 4/(\sqrt{L} + \sqrt{\mu}), \quad \beta_k = q^2, q = \frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1}$$

the heavy-ball method satisfies $\|x^k - x^*\| = O(q^k)$.

Motivated originally from so-called "Ravine method" of Gelfand-Tsetlin (1961), that runs the iteration

$$z^k = x^k - \eta_k \nabla f(x^k), \quad x^{k+1} = z^k + \beta_k(z^k - z^{k-1})$$

Background: Heavy-ball method

Polyak's Momentum Method (1964)

$$x^{k+1} = x^k - \eta_k \nabla f(x^k) + \beta_k (x^k - x^{k-1})$$

Can view it as a discretization of 2nd-order ODE:

$$\ddot{x} + a\dot{x} + b\nabla f(x) = 0$$

(analogy: movement of a heavy-ball in a potential field $f(x)$ governed not only by $\nabla f(x)$ but by a ***momentum*** term)

Background: Heavy-ball method

Polyak's Momentum Method (1964)

$$x^{k+1} = x^k - \eta_k \nabla f(x^k) + \beta_k (x^k - x^{k-1})$$

Can view it as a discretization of 2nd-order ODE:

$$\ddot{x} + a\dot{x} + b\nabla f(x) = 0$$

(analogy: movement of a heavy-ball in a potential field $f(x)$ governed not only by $\nabla f(x)$ but by a ***momentum*** term)

Why does momentum help?

Explore: Check out: *https://distill.pub/2017/momentum/*

What about the general convex case?

Nesterov's AGM

Nesterov's AGM

Nesterov's (1983) method

$$
\begin{aligned}
x^{k+1} &= y^k - \tfrac{1}{L}\nabla f(y^k) \\
y^{k+1} &= x^{k+1} + \beta_k(x^{k+1} - x^k)
\end{aligned}
$$

Nesterov's AGM

Nesterov's (1983) method

$$\begin{aligned} x^{k+1} &= y^k - \tfrac{1}{L}\nabla f(y^k) \\ y^{k+1} &= x^{k+1} + \beta_k(x^{k+1} - x^k) \end{aligned}$$

Essentially same as the ravine method!!

Nesterov's AGM

Nesterov's (1983) method

$$\begin{aligned} x^{k+1} &= y^k - \tfrac{1}{L}\nabla f(y^k) \\ y^{k+1} &= x^{k+1} + \beta_k(x^{k+1} - x^k) \end{aligned}$$

Essentially same as the ravine method!!

$$\beta_k = \frac{\alpha_k - 1}{\alpha_{k+1}}, \qquad 2\alpha_{k+1} = 1 + \sqrt{4\alpha_k^2 + 1},\ \alpha_0 = 1$$

$$f(x^k) - f(x^*) \le \frac{2L\|y_0 - x^*\|^2}{(k+2)^2}.$$

In the strongly convex case, instead we use $\beta_k = \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}$. This leads to $O(\sqrt{\kappa}\log(1/\varepsilon))$ iterations to ensure $f(x^k) - f(x^*) \le \varepsilon$.

(**Remark**: Nemirovski proposed a method that achieves optimal complexity, but it required 2D line-search. Nesterov's method was the real breakthrough and remains a fascinating topic to study even today.)

Analyzing Nesterov's method

(►► Ravine method worked well and sparked numerous heuristics for selecting its parameters and improving its behavior. However, its convergence was never proved. Inspired Polyak's heavy-ball method, which seems to have inspired Nesterov's AGM.)

Analyzing Nesterov's method

(►► Ravine method worked well and sparked numerous heuristics for selecting its parameters and improving its behavior. However, its convergence was never proved. Inspired Polyak's heavy-ball method, which seems to have inspired Nesterov's AGM.)

Some ways to analyze AGM

- Nesterov's Estimate sequence method
- Approaches based on potential (Lyapunov) functions
- Derivation based on viewing AGM as approximate PPM
- Using "linear coupling," mixing a primal-dual view
- Analysis based on SDPs

Analyzing Nesterov's method

(►► Ravine method worked well and sparked numerous heuristics for selecting its parameters and improving its behavior. However, its convergence was never proved. Inspired Polyak's heavy-ball method, which seems to have inspired Nesterov's AGM.)

Some ways to analyze AGM

- Nesterov's Estimate sequence method
- Approaches based on potential (Lyapunov) functions
- Derivation based on viewing AGM as approximate PPM
- Using "linear coupling," mixing a primal-dual view
- Analysis based on SDPs

See discussion in the paper

From Nesterov's Estimate Sequence to Riemannian Acceleration

Kwangjun Ahn KJAHN@MIT.EDU

Suvrit Sra SUVRIT@MIT.EDU

Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology

Potential analysis – sketch

- Choose potential: judge closeness of iterates to the optimal
- Ensure the potential is decreasing with iteration
- AGM does not satisfy $f(x^{k+1}) \leq f(x^k)$, so...

Potential analysis – sketch

- Choose potential: judge closeness of iterates to the optimal
- Ensure the potential is decreasing with iteration
- AGM does not satisfy $f(x^{k+1}) \leq f(x^k)$, so...

Slightly more general AGM iteration

$$
\begin{aligned}
x^{k+1} &\leftarrow y^k + \alpha_{k+1}(z^k - y^k) \\
y^{k+1} &\leftarrow x^{k+1} - \gamma_{k+1}\nabla f(x^{k+1}) \\
z^{k+1} &\leftarrow x^{k+1} + \beta_{k+1}(z^k - x^{k+1}) - \eta_{k+1}\nabla f(x^{k+1})
\end{aligned}
$$

Mixing intution from "descent" and "ravines"

$$\Phi_k := A_k(f(y^k) - f(x^*)) + B_k\|z^k - x^*\|^2$$

Pick parameters $A_k, B_k, \eta_k, \gamma_k, \alpha_k, \beta_k$ to ensure that we have $\Phi_k - \Phi_{k-1} \leq 0$. Turns out a "simple" choice does that job!

Potential analysis – sketch

Using the shorthand:

$$\Delta_\gamma := \gamma(1 - L\gamma/2)\,, \quad \nabla := \nabla f(x_{t+1})\,, \quad X := x_{t+1} - x_*\,, \text{ and } W := z_t - x_{t+1},$$

using smoothness and convexity, show that $\Phi_{k+1} - \Phi_k$ is upper-boudned by

$$c_1\|W\|^2 + c_2\|X\|^2 + c_3\|\nabla\|^2 + c_4\langle W, X\rangle + c_5\langle W, \nabla\rangle + c_6\langle X, \nabla\rangle\,,$$

$$\begin{cases} c_1 := \beta^2 B_{k+1} - B_k - \frac{\mu}{2}\frac{\alpha^2}{(1-\alpha)^2}A_k\,, & c_2 := B_{k+1} - B_k - \frac{\mu}{2}(A_{k+1} - A_k)\,, \\ c_3 := \eta^2 B_{k+1} - \Delta_\gamma \cdot A_{k+1}\,, & c_4 := 2\cdot(\beta B_{k+1} - B_k)\,, \\ c_5 := \frac{\alpha}{1-\alpha}A_k - 2\beta\eta B_{k+1}\,, \quad \text{and} & c_6 := (A_{k+1} - A_k) - 2\eta B_{k+1}\,. \end{cases}$$

Potential analysis – sketch

Using the shorthand:

$$\Delta_\gamma := \gamma(1 - L\gamma/2)\,, \quad \nabla := \nabla f(x_{t+1})\,, \quad X := x_{t+1} - x_*\,, \text{ and } W := z_t - x_{t+1},$$

using smoothness and convexity, show that $\Phi_{k+1} - \Phi_k$ is upper-boudned by

$$c_1\|W\|^2 + c_2\|X\|^2 + c_3\|\nabla\|^2 + c_4\langle W, X\rangle + c_5\langle W, \nabla\rangle + c_6\langle X, \nabla\rangle\,,$$

$$\begin{cases} c_1 := \beta^2 B_{k+1} - B_k - \frac{\mu}{2}\frac{\alpha^2}{(1-\alpha)^2}A_k\,, & c_2 := B_{k+1} - B_k - \frac{\mu}{2}(A_{k+1} - A_k)\,, \\ c_3 := \eta^2 B_{k+1} - \Delta_\gamma \cdot A_{k+1}\,, & c_4 := 2\cdot(\beta B_{k+1} - B_k)\,, \\ c_5 := \frac{\alpha}{1-\alpha}A_k - 2\beta\eta B_{k+1}\,, \quad \text{and} & c_6 := (A_{k+1} - A_k) - 2\eta B_{k+1}\,. \end{cases}$$

Now choose parameters to ensure $\Phi_{k+1} - \Phi_k \leq 0$. Finally, leads to a bound of the form

$$f(y^k) - f(x^*) = O((1 - \xi_1)\cdots(1 - \xi_k)),$$

where the sequence $\{\xi_k\}$ fully characterizes convergence.

Ref: See details in the paper: Ahn, Sra (2020). *From Nesterov's Estimate Sequence to Riemannian Acceleration*.