Subgradient methods

Yuanming Shi and Ye Shi

ShanghaiTech University

Copyright: ELE 522: "Large-Scale Optimization for Data Science" from Prof. Yuxin Chen at Princeton University

Outline

- Steepest descent
- Subgradients
- Projected subgradient descent
  - Convex and Lipschitz problems
  - Strongly convex and Lipschitz problems
- Convex-concave saddle point problems

Nondifferentiable problems

Differentiability of the objective function f is essential for the validity of gradient methods

However, there is no shortage of interesting cases (e.g. ℓ_1 minimization, nuclear norm minimization) where non-differentiability is present at some points

Generalizing steepest descent?

$$\text{minimize}_{\boldsymbol{x}} \ f(\boldsymbol{x}) \qquad \text{subject to } \boldsymbol{x} \in \mathcal{C}$$

- find a search direction $\boldsymbol{d}^t$ that minimizes the directional derivative

$$\boldsymbol{d}^t \in \arg\min_{\boldsymbol{d}:\|\boldsymbol{d}\|_2 \leq 1} f'(\boldsymbol{x}^t; \boldsymbol{d})$$

where $f'(\boldsymbol{x}; \boldsymbol{d}) := \lim_{\alpha \downarrow 0} \frac{f(\boldsymbol{x}+\alpha \boldsymbol{d}) - f(\boldsymbol{x})}{\alpha}$

- updates

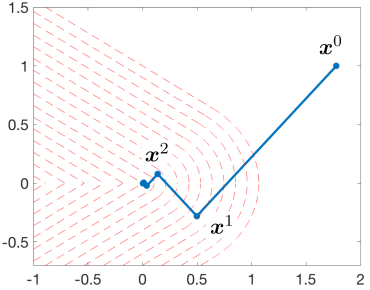
$$\boldsymbol{x}^{t+1} = \boldsymbol{x}^t + \eta_t \boldsymbol{d}^t$$

Issues

- Finding the steepest descent direction (or even finding a descent direction) may involve expensive computation

- Stepsize rules are tricky to choose: for certain popular stepsize rules (like exact line search), steepest descent might converge to non-optimal points

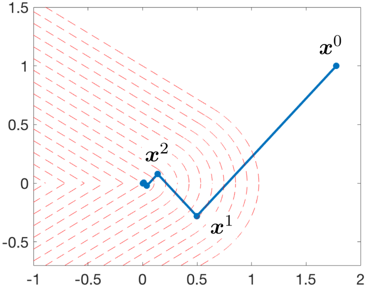
Wolfe's example



$$f(x_1, x_2) = \begin{cases} 5(9x_1^2 + 16x_2^2)^{\frac{1}{2}} & \text{if } x_1 > |x_2| \\ 9x_1 + 16|x_2| & \text{if } x_1 \leq |x_2| \end{cases}$$

- (0,0) is a non-differentiable point
- if one starts from $\boldsymbol{x}^0 = (\frac{16}{9}, 1)$ and uses exact line search, then
  - $\{\boldsymbol{x}^t\}$ are all differentiable points
  - $\boldsymbol{x}^t \to (0,0)$ as $t \to \infty$

Wolfe's example



$$f(x_1, x_2) = \begin{cases} 5(9x_1^2 + 16x_2^2)^{\frac{1}{2}} & \text{if } x_1 > |x_2| \\ 9x_1 + 16|x_2| & \text{if } x_1 \le |x_2| \end{cases}$$

- even though it never hits non-differentiable points, steepest descent with exact line search gets stuck around a non-optimal point (i.e. $(0, 0)$)
- **problem**: steepest descent directions may undergo large / discontinuous changes when close to convergence limits

(Projected) subgradient method

Practically, a popular choice is "subgradient-based methods"

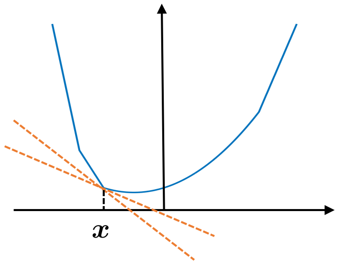
$$\boldsymbol{x}^{t+1} = \mathcal{P}_{\mathcal{C}}\left(\boldsymbol{x}^t - \eta_t \boldsymbol{g}^t\right) \tag{4.1}$$

where $\boldsymbol{g}^t$ is any subgradient of f at $\boldsymbol{x}^t$

- the focus of this lecture
- **caution:** this update rule does not necessarily yield reduction w.r.t. the objective values

Subgradients

Subgradients

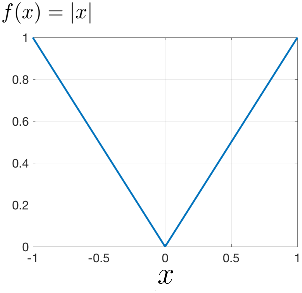


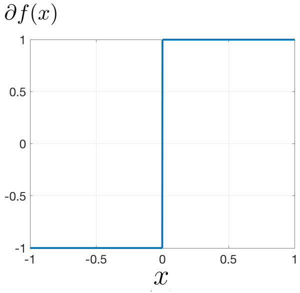
We say $\boldsymbol{g}$ is a subgradient of f at the point $\boldsymbol{x}$ if

$$f(\boldsymbol{z}) \geq \underbrace{f(\boldsymbol{x}) + \boldsymbol{g}^\top(\boldsymbol{z} - \boldsymbol{x})}_{\text{a linear } \textit{under}\text{-estimate of } f}, \quad \forall \boldsymbol{z} \tag{4.2}$$

- the set of all subgradients of f at $\boldsymbol{x}$ is called the subdifferential of f at $\boldsymbol{x}$, denoted by $\partial f(\boldsymbol{x})$

Example: $f(x) = |x|$





$$f(x) = |x| \qquad\qquad \partial f(x) = \begin{cases} \{-1\}, & \text{if } x < 0 \\ [-1, 1], & \text{if } x = 0 \\ \{1\}, & \text{if } x > 0 \end{cases}$$

Example: a subgradient of norms at 0

Let $f(\boldsymbol{x}) = \|\boldsymbol{x}\|$ for any norm $\|\cdot\|$, then for any $\boldsymbol{g}$ obeying $\|\boldsymbol{g}\|_* \leq 1$,

$$\boldsymbol{g} \in \partial f(\boldsymbol{0})$$

where $\|\cdot\|_*$ is the dual norm of $\|\cdot\|$ (i.e. $\|\boldsymbol{x}\|_* := \sup_{\boldsymbol{z}:\|\boldsymbol{z}\|\leq 1}\langle \boldsymbol{z}, \boldsymbol{x}\rangle$)

Proof: To see this, it suffices to prove that

$$f(\boldsymbol{z}) \geq f(\boldsymbol{0}) + \langle \boldsymbol{g}, \boldsymbol{z} - \boldsymbol{0}\rangle, \qquad \forall \boldsymbol{z}$$

$$\Longleftrightarrow \quad \langle \boldsymbol{g}, \boldsymbol{z}\rangle \leq \|\boldsymbol{z}\|, \qquad \forall \boldsymbol{z}$$

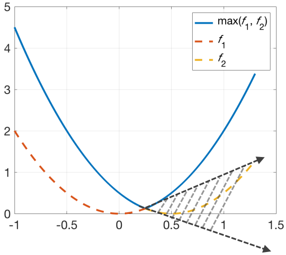
This follows from generalized Cauchy-Schwarz, i.e.

$$\langle \boldsymbol{g}, \boldsymbol{z}\rangle \leq \|\boldsymbol{g}\|_* \|\boldsymbol{z}\| \leq \|\boldsymbol{z}\|$$

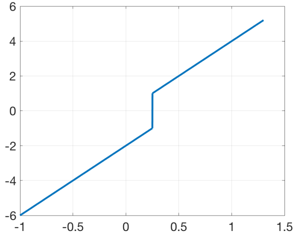
$\square$

Example: $\max\{f_1(x), f_2(x)\}$

$f(x) = \max\{f_1(x), f_2(x)\}$



$\partial f(x)$



$f(x) = \max\{f_1(x), f_2(x)\}$ where f_1 and f_2 are differentiable

$$\partial f(x) = \begin{cases} \{f_1'(x)\}, & \text{if } f_1(x) > f_2(x) \\ [f_1'(x), f_2'(x)], & \text{if } f_1(x) = f_2(x) \\ \{f_2'(x)\}, & \text{if } f_1(x) < f_2(x) \end{cases}$$

Basic rules

- **scaling:** $\partial(\alpha f) = \alpha \partial f$ (for $\alpha > 0$)
- **summation:** $\partial(f_1 + f_2) = \partial f_1 + \partial f_2$

Example: ℓ_1 norm

$$f(\boldsymbol{x}) = \|\boldsymbol{x}\|_1 = \sum_{i=1}^{n} \underbrace{|x_i|}_{=: f_i(\boldsymbol{x})}$$

since

$$\partial f_i(\boldsymbol{x}) = \begin{cases} \operatorname{sgn}(x_i)\boldsymbol{e}_i, & \text{if } x_i \neq 0 \\ [-1,1]\cdot \boldsymbol{e}_i, & \text{if } x_i = 0 \end{cases}$$

we have

$$\sum_{i: x_i \neq 0} \operatorname{sgn}(x_i)\boldsymbol{e}_i \in \partial f(\boldsymbol{x})$$

Basic rules (cont.)

- **affine transformation:** if $h(\boldsymbol{x}) = f(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b})$, then

$$\partial h(\boldsymbol{x}) = \boldsymbol{A}^{\top} \partial f(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b})$$

Example: $\|\boldsymbol{A}\boldsymbol{x}+\boldsymbol{b}\|_1$

$$h(\boldsymbol{x}) = \|\boldsymbol{A}\boldsymbol{x}+\boldsymbol{b}\|_1$$

letting $f(\boldsymbol{x}) = \|\boldsymbol{x}\|_1$ and $\boldsymbol{A} = [\boldsymbol{a}_1, \cdots, \boldsymbol{a}_m]^\top$, we have

$$\boldsymbol{g} = \sum_{i:\boldsymbol{a}_i^\top \boldsymbol{x}+b_i \neq 0} \mathrm{sgn}(\boldsymbol{a}_i^\top \boldsymbol{x}+b_i)\boldsymbol{e}_i \ \in\ \partial f(\boldsymbol{A}\boldsymbol{x}+\boldsymbol{b}).$$

$$\Longrightarrow \quad \boldsymbol{A}^\top \boldsymbol{g} = \sum_{i:\boldsymbol{a}_i^\top \boldsymbol{x}+b_i \neq 0} \mathrm{sgn}(\boldsymbol{a}_i^\top \boldsymbol{x}+b_i)\boldsymbol{a}_i \ \in\ \partial h(\boldsymbol{x})$$

Basic rules (cont.)

- **chain rule:** suppose f is convex, and g is differentiable, nondecreasing, and convex. Let $h = g \circ f$, then

$$\partial h(\boldsymbol{x}) = g'(f(\boldsymbol{x}))\partial f(\boldsymbol{x})$$

- **composition:** suppose $f(\boldsymbol{x}) = h(f_1(\boldsymbol{x}), \cdots, f_n(\boldsymbol{x}))$, where f_i's are convex, and h is differentiable, nondecreasing, and convex. Let $\boldsymbol{q} = \nabla h\left(\boldsymbol{y}\right)\big|_{\boldsymbol{y}=[f_1(\boldsymbol{x}),\cdots,f_n(\boldsymbol{x})]}$, and $\boldsymbol{g}_i \in \partial f_i(\boldsymbol{x})$. Then

$$q_1\boldsymbol{g}_1 + \cdots + q_n\boldsymbol{g}_n \ \in \ \partial f(\boldsymbol{x})$$

Basic rules (cont.)

- **pointwise maximum:** if $f(\boldsymbol{x}) = \max_{1\leq i\leq k} f_i(\boldsymbol{x})$, then

$$\partial f(\boldsymbol{x}) = \underbrace{\mathsf{conv}\left\{\bigcup \left\{\partial f_i(\boldsymbol{x}) \mid f_i(\boldsymbol{x}) = f(\boldsymbol{x})\right\}\right\}}_{\text{convex hull of subdifferentials of all active functions}}$$

- **pointwise supremum:** if $f(\boldsymbol{x}) = \sup_{\alpha\in\mathcal{F}} f_\alpha(\boldsymbol{x})$, then

$$\partial f(\boldsymbol{x}) = \mathsf{closure}\left(\mathsf{conv}\left\{\bigcup \left\{\partial f_\alpha(\boldsymbol{x}) \mid f_\alpha(\boldsymbol{x}) = f(\boldsymbol{x})\right\}\right\}\right)$$

Example: piece-wise linear functions

$$f(\boldsymbol{x}) = \max_{1 \leq i \leq m} \left\{ \boldsymbol{a}_i^\top \boldsymbol{x} + b_i \right\}$$

pick any $\boldsymbol{a}_j$ s.t. $\boldsymbol{a}_j^\top \boldsymbol{x} + b_j = \max_i \left\{ \boldsymbol{a}_i^\top \boldsymbol{x} + b_i \right\}$, then

$$\boldsymbol{a}_j \in \partial f(\boldsymbol{x})$$

Example: the ℓ_∞ norm

$$f(\boldsymbol{x}) = \|\boldsymbol{x}\|_\infty = \max_{1\leq i\leq n} |x_i|$$

if $\boldsymbol{x} \neq \mathbf{0}$, then pick any x_j obeying $|x_j| = \max_i |x_i|$ to obtain

$$\mathrm{sgn}(x_j)\boldsymbol{e}_j \in \partial f(\boldsymbol{x})$$

Example: the maximum eigenvalue

$$f(\boldsymbol{x}) = \lambda_{\max}\left(x_1 \boldsymbol{A}_1 + \cdots + x_n \boldsymbol{A}_n\right)$$

where $\boldsymbol{A}_1, \cdots, \boldsymbol{A}_n$ are real symmetric matrices

Rewrite

$$f(\boldsymbol{x}) = \sup_{\boldsymbol{y}:\|\boldsymbol{y}\|_2=1} \boldsymbol{y}^\top \left(x_1 \boldsymbol{A}_1 + \cdots + x_n \boldsymbol{A}_n\right) \boldsymbol{y}$$

as the supremum of some affine functions of $\boldsymbol{x}$. Therefore, taking $\boldsymbol{y}$ as the leading eigenvector of $x_1 \boldsymbol{A}_1 + \cdots + x_n \boldsymbol{A}_n$, we have

$$\left[\boldsymbol{y}^\top \boldsymbol{A}_1 \boldsymbol{y}, \cdots, \boldsymbol{y}^\top \boldsymbol{A}_n \boldsymbol{y}\right]^\top \in \partial f(\boldsymbol{x})$$

Example: the nuclear norm

Let $\boldsymbol{X} \in \mathbb{R}^{m \times n}$ with SVD $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$ and

$$f(\boldsymbol{X}) = \sum_{i=1}^{\min\{n,m\}} \sigma_i(\boldsymbol{X})$$

where $\sigma_i(\boldsymbol{x})$ is the ith largest singular value of $\boldsymbol{X}$

Rewrite

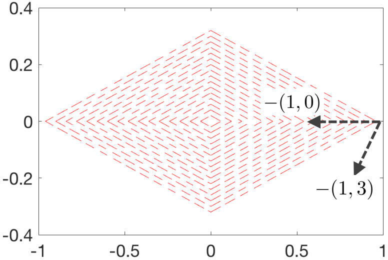
$$f(\boldsymbol{X}) = \sup_{\text{orthonormal } \boldsymbol{A},\boldsymbol{B}} \langle \boldsymbol{A}\boldsymbol{B}^\top, \boldsymbol{X} \rangle := \sup_{\text{orthonormal } \boldsymbol{A},\boldsymbol{B}} f_{\boldsymbol{A},\boldsymbol{B}}(\boldsymbol{X})$$

Recognizing that $f_{\boldsymbol{A},\boldsymbol{B}}(\boldsymbol{X})$ is maximized by $\boldsymbol{A} = \boldsymbol{U}$ and $\boldsymbol{B} = \boldsymbol{V}$ and that $\nabla f_{\boldsymbol{A},\boldsymbol{B}}(\boldsymbol{X}) = \boldsymbol{A}\boldsymbol{B}^\top$, we have

$$\boldsymbol{U}\boldsymbol{V}^\top \in \partial f(\boldsymbol{X})$$

Negative subgradients are not necessarily descent directions

Example: $f(\boldsymbol{x}) = |x_1| + 3|x_2|$



at $\boldsymbol{x} = (1, 0)$:

- $\boldsymbol{g}_1 = (1, 0) \in \partial f(\boldsymbol{x})$, and $-\boldsymbol{g}_1$ is a descent direction
- $\boldsymbol{g}_2 = (1, 3) \in \partial f(\boldsymbol{x})$, but $-\boldsymbol{g}_2$ is not a descent direction

Reason: lack of continuity — one can change directions significantly without violating the validity of subgradients

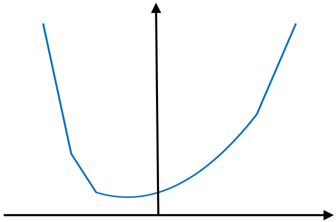
Negative subgradient is not necessarily descent direction

Since $f(\boldsymbol{x}^t)$ is not necessarily monotone, we will keep track of the best point

$$f^{\text{best},t} := \min_{1 \leq i \leq t} f(\boldsymbol{x}^i)$$

We also denote by $f^{\text{opt}} := \min_{\boldsymbol{x}} f(\boldsymbol{x})$ the optimal objective value

Convex and Lipschitz problems



Clearly, we cannot analyze all nonsmooth functions. A nice (and widely encountered) class to start with is Lipschitz functions, i.e. the set of all f obeying

$$|f(\boldsymbol{x}) - f(\boldsymbol{z})| \leq L_f \|\boldsymbol{x} - \boldsymbol{z}\|_2 \qquad \forall\, \boldsymbol{x} \text{ and } \boldsymbol{z}$$

Fundamental inequality for projected subgradient methods

We'd like to optimize $\|\boldsymbol{x}^{t+1} - \boldsymbol{x}^*\|_2^2$, but don't have access to $\boldsymbol{x}^*$

Key idea (majorization-minimization): find another function that majorizes $\|\boldsymbol{x}^{t+1} - \boldsymbol{x}^*\|_2^2$, and optimize the majorizing function

Lemma 4.1

Projected subgradient update rule (4.1) obeys

$$\|\boldsymbol{x}^{t+1} - \boldsymbol{x}^*\|_2^2 \leq \underbrace{\underbrace{\|\boldsymbol{x}^t - \boldsymbol{x}^*\|_2^2}_{\textit{fixed}} - 2\eta_t\big(f(\boldsymbol{x}^t) - f^{\mathsf{opt}}\big) + \eta_t^2\|\boldsymbol{g}^t\|_2^2}_{\textit{majorizing function}} \quad (4.3)$$

Proof of Lemma 4.1

$$\begin{aligned}
\|\boldsymbol{x}^{t+1} - \boldsymbol{x}^*\|_2^2 &= \|\mathcal{P}_{\mathcal{C}}(\boldsymbol{x}^t - \eta_t \boldsymbol{g}^t) - \mathcal{P}_{\mathcal{C}}(\boldsymbol{x}^*)\|_2^2 \\
&\leq \|\boldsymbol{x}^t - \eta_t \boldsymbol{g}^t - \boldsymbol{x}^*\|_2^2 \qquad \text{(nonexpansiveness of projection)} \\
&= \|\boldsymbol{x}^t - \boldsymbol{x}^*\|_2^2 - 2\eta_t \langle \boldsymbol{x}^t - \boldsymbol{x}^*, \boldsymbol{g}^t \rangle + \eta_t^2 \|\boldsymbol{g}^t\|_2^2 \\
&\leq \|\boldsymbol{x}^t - \boldsymbol{x}^*\|_2^2 - 2\eta_t \big(f(\boldsymbol{x}^t) - f(\boldsymbol{x}^*)\big) + \eta_t^2 \|\boldsymbol{g}^t\|_2^2
\end{aligned}$$

where the last line uses the subgradient inequality

$$f(\boldsymbol{x}^*) - f(\boldsymbol{x}^t) \geq \langle \boldsymbol{x}^* - \boldsymbol{x}^t, \boldsymbol{g}^t \rangle$$

Polyak's stepsize rule

The majorizing function in (4.3) suggests a stepsize (Polyak '87)

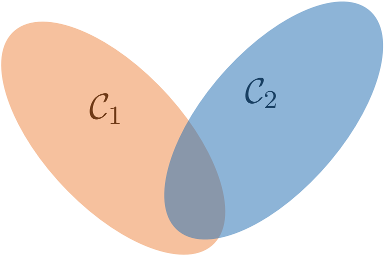
$$\eta_t = \frac{f(\boldsymbol{x}^t) - f^{\mathsf{opt}}}{\|\boldsymbol{g}_t\|_2^2} \tag{4.4}$$

which leads to error reduction

$$\|\boldsymbol{x}^{t+1} - \boldsymbol{x}^*\|_2^2 \le \|\boldsymbol{x}^t - \boldsymbol{x}^*\|_2^2 - \frac{\big(f(\boldsymbol{x}^t) - f(\boldsymbol{x}^*)\big)^2}{\|\boldsymbol{g}^t\|_2^2} \tag{4.5}$$

- useful if f^{opt} is known
- the estimation error is monotonically decreasing with Polyak's stepsize

Example: projection onto intersection of convex sets



Let $\mathcal{C}_1$, $\mathcal{C}_2$ be closed convex sets and suppose $\mathcal{C}_1 \cap \mathcal{C}_2 \neq \emptyset$

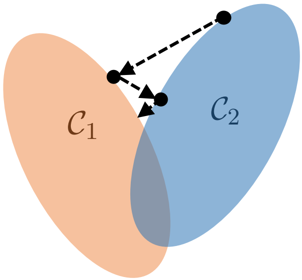
$$\text{find} \quad \boldsymbol{x} \in \mathcal{C}_1 \cap \mathcal{C}_2$$

$$\Updownarrow$$

$$\text{minimize}_{\boldsymbol{x}} \quad \max\left\{\text{dist}_{\mathcal{C}_1}(\boldsymbol{x}), \text{dist}_{\mathcal{C}_2}(\boldsymbol{x})\right\}$$

where $\text{dist}_{\mathcal{C}}(\boldsymbol{x}) := \min_{\boldsymbol{z} \in \mathcal{C}} \|\boldsymbol{x} - \boldsymbol{z}\|_2$

Example: projection onto intersection of convex sets



For this problem, the subgradient method with Polyak's stepsize rule is equivalent to alternating projection

$$\boldsymbol{x}^{t+1} = \mathcal{P}_{\mathcal{C}_1}(\boldsymbol{x}^t), \quad \boldsymbol{x}^{t+2} = \mathcal{P}_{\mathcal{C}_2}(\boldsymbol{x}^{t+1})$$

Example: projection onto intersection of convex sets

Proof: Use the subgradient rule for pointwise max functions to get

$$\boldsymbol{g}^t \in \partial \mathsf{dist}_{\mathcal{C}_i}(\boldsymbol{x}^t)$$

where $i = \arg\max_{j=1,2} \mathsf{dist}_{\mathcal{C}_j}(\boldsymbol{x}^t)$

If $\mathsf{dist}_{\mathcal{C}_i}(\boldsymbol{x}^t) \neq 0$, then one has

$$\boldsymbol{g}^t = \nabla \mathsf{dist}_{\mathcal{C}_i}(\boldsymbol{x}^t) = \frac{\boldsymbol{x}^t - \mathcal{P}_{\mathcal{C}_i}(\boldsymbol{x}^t)}{\mathsf{dist}_{C_i}(\boldsymbol{x}^t)}$$

which follows since $\nabla\Big(\frac{1}{2}\mathsf{dist}^2_{\mathcal{C}_i}(\boldsymbol{x}^t)\Big) = \boldsymbol{x}^t - \mathcal{P}_{\mathcal{C}_i}(\boldsymbol{x}^t)$ (homework) and $\nabla\Big(\frac{1}{2}\mathsf{dist}^2_{\mathcal{C}_i}(\boldsymbol{x}^t)\Big) = \mathsf{dist}_{\mathcal{C}_i}(\boldsymbol{x}^t) \cdot \nabla \mathsf{dist}_{\mathcal{C}_i}(\boldsymbol{x}^t)$

Example: projection onto intersection of convex sets

Proof (cont.): Adopting Polya's stepsize rule and recognizing that $\|\boldsymbol{g}^t\|_2 = 1$, we arrive at

$$\begin{aligned}
\boldsymbol{x}^{t+1} = \boldsymbol{x}^t - \eta_t \boldsymbol{g}^t &= \boldsymbol{x}^t - \underbrace{\frac{\mathsf{dist}_{\mathcal{C}_i}(\boldsymbol{x}^t)}{\|\boldsymbol{g}^t\|_2^2}}_{=\eta_t} \frac{\boldsymbol{x}^t - \mathcal{P}_{\mathcal{C}_i}(\boldsymbol{x}^t)}{\mathsf{dist}_{C_i}(\boldsymbol{x}^t)} \\
&= \mathcal{P}_{\mathcal{C}_i}(\boldsymbol{x}^t)
\end{aligned}$$

where $i = \arg\max_{j=1,2} \mathsf{dist}_{\mathcal{C}_j}(\boldsymbol{x}^t)$

$\square$

Convergence rate with Polyak's stepsize

Theorem 4.2 (Convergence of projected subgradient method with Polyak's stepsize)

Suppose f is convex and L_f-Lipschitz continuous. Then the projected subgradient method (4.1) with Polyak's stepsize rule obeys

$$f^{\mathsf{best},t} - f^{\mathsf{opt}} \leq \frac{L_f \|\boldsymbol{x}^0 - \boldsymbol{x}^*\|_2}{\sqrt{t+1}}$$

- sublinear convergence rate $O(1/\sqrt{t})$

Proof of Theorem 4.2

We have seen from (4.5) that

$$\begin{aligned}\big(f(\boldsymbol{x}^t) - f(\boldsymbol{x}^*)\big)^2 &\le \Big\{\|\boldsymbol{x}^t - \boldsymbol{x}^*\|_2^2 - \|\boldsymbol{x}^{t+1} - \boldsymbol{x}^*\|_2^2\Big\}\|\boldsymbol{g}^t\|_2^2 \\ &\le \Big\{\|\boldsymbol{x}^t - \boldsymbol{x}^*\|_2^2 - \|\boldsymbol{x}^{t+1} - \boldsymbol{x}^*\|_2^2\Big\} L_f^2\end{aligned}$$

Applying it recursively for all iterations (from 0th to tth) and summing them up yield

$$\sum_{k=0}^{t}\big(f(\boldsymbol{x}^k) - f(\boldsymbol{x}^*)\big)^2 \le \Big\{\|\boldsymbol{x}^0 - \boldsymbol{x}^*\|_2^2 - \|\boldsymbol{x}^{t+1} - \boldsymbol{x}^*\|_2^2\Big\} L_f^2$$

$$\Longrightarrow \quad (t+1)\big(f^{\mathsf{best},t} - f^{\mathsf{opt}}\big)^2 \le \|\boldsymbol{x}^0 - \boldsymbol{x}^*\|_2^2 L_f^2$$

which concludes the proof

Other stepsize choices?

Unfortunately, Polyak's stepsize rule requires knowledge of f^{opt}, which is often unknown *a priori*

We might often need simpler rules for setting stepsizes

Convex and Lipschitz problems

Theorem 4.3 (Subgradient methods for convex and Lipschitz functions)

Suppose f is convex and L_f-Lipschitz continuous. Then the projected subgradient update rule (4.1) obeys

$$f^{\mathsf{best},t} - f^{\mathsf{opt}} \leq \frac{\|\boldsymbol{x}^0 - \boldsymbol{x}^*\|_2^2 + L_f^2 \sum_{i=0}^{t} \eta_i^2}{2 \sum_{i=0}^{t} \eta_i}$$

Implications: stepsize rules

- **Constant step size** $\eta_t \equiv \eta$:

$$\lim_{t\to\infty} f^{\mathsf{best},t} \leq \frac{L_f^2 \eta}{2}$$

  i.e. may converge to non-optimal points

- **Diminishing step size obeying** $\sum_t \eta_t^2 < \infty$ **and** $\sum_t \eta_t \to \infty$:

$$\lim_{t\to\infty} f^{\mathsf{best},t} = 0$$

  i.e. converges to optimal points

Implications: stepsize rule

- **Optimal choice?** $\eta_t = \frac{1}{\sqrt{t}}$:

$$f^{\mathsf{best},t} - f^{\mathsf{opt}} \lesssim \frac{\|\boldsymbol{x}^0 - \boldsymbol{x}^*\|_2^2 + L_f^2 \log t}{\sqrt{t}}$$

  i.e. attains ε-accuracy within about $O(1/\varepsilon^2)$ iterations (ignoring the log factor)

Proof of Theorem 4.5

Applying Lemma 4.1 recursively gives

$$\|\boldsymbol{x}^{t+1} - \boldsymbol{x}^*\|_2^2 \leq \|\boldsymbol{x}^0 - \boldsymbol{x}^*\|_2^2 - 2\sum_{i=0}^{t} \eta_i\big(f(\boldsymbol{x}^i) - f^{\mathsf{opt}}\big) + \sum_{i=0}^{t} \eta_i^2 \|\boldsymbol{g}^i\|_2^2$$

Rearranging terms, we are left with

$$\begin{aligned} 2\sum_{i=0}^{t} \eta_i\big(f(\boldsymbol{x}^i) - f^{\mathsf{opt}}\big) &\leq \|\boldsymbol{x}^0 - \boldsymbol{x}^*\|_2^2 - \|\boldsymbol{x}^{t+1} - \boldsymbol{x}^*\|_2^2 + \sum_{i=0}^{t} \eta_i^2 \|\boldsymbol{g}^i\|_2^2 \\ &\leq \|\boldsymbol{x}^0 - \boldsymbol{x}^*\|_2^2 + L_f^2 \sum_{i=0}^{t} \eta_i^2 \end{aligned}$$

$$\Longrightarrow \quad f^{\mathsf{best},t} - f^{\mathsf{opt}} \leq \frac{\sum_{i=0}^{t} \eta_i\big(f(\boldsymbol{x}^i) - f^{\mathsf{opt}}\big)}{\sum_{i=0}^{t} \eta_i} \leq \frac{\|\boldsymbol{x}^0 - \boldsymbol{x}^*\|_2^2 + L_f^2 \sum_{i=0}^{t} \eta_i^2}{2\sum_{i=0}^{t} \eta_i}$$

Strongly convex and Lipschitz problems

If f is strongly convex, then the convergence guarantees can be improved to $O(1/t)$, as long as the stepsize dimishes at $O(1/t)$

Theorem 4.4 (Subgradient methods for strongly convex and Lipschitz functions)

Let f be μ-strongly convex and L_f-Lipschitz continuous over $\mathcal{C}$. If $\eta_t \equiv \eta = \frac{2}{\mu(t+1)}$, then

$$f^{\mathsf{best},t} - f^{\mathsf{opt}} \leq \frac{2L_f^2}{\mu} \cdot \frac{1}{t+1}$$

Proof of Theorem 4.4

When f is μ-strongly convex, we can improve Lemma 4.1 to (exercise)

$$\|\boldsymbol{x}^{t+1} - \boldsymbol{x}^*\|_2^2 \leq (1 - \mu\eta_t)\|\boldsymbol{x}^t - \boldsymbol{x}^*\|_2^2 - 2\eta_t \left(f(\boldsymbol{x}^t) - f^{\mathrm{opt}}\right) + \eta_t^2\|\boldsymbol{g}^t\|_2^2$$

$$\Longrightarrow \quad f(\boldsymbol{x}^t) - f^{\mathrm{opt}} \leq \frac{1 - \mu\eta_t}{2\eta_t}\|\boldsymbol{x}^t - \boldsymbol{x}^*\|_2^2 - \frac{1}{2\eta_t}\|\boldsymbol{x}^{t+1} - \boldsymbol{x}^*\|_2^2 + \frac{\eta_t}{2}\|\boldsymbol{g}^t\|_2^2$$

Since $\eta_t = 2/(\mu(t+1))$, we have

$$f(\boldsymbol{x}^t) - f^{\mathrm{opt}} \leq \frac{\mu(t-1)}{4}\|\boldsymbol{x}^t - \boldsymbol{x}^*\|_2^2 - \frac{\mu(t+1)}{4}\|\boldsymbol{x}^{t+1} - \boldsymbol{x}^*\|_2^2 + \frac{1}{\mu(t+1)}\|\boldsymbol{g}^t\|_2^2$$

and hence

$$t\left(f(\boldsymbol{x}^t) - f^{\mathrm{opt}}\right) \leq \frac{\mu t(t-1)}{4}\|\boldsymbol{x}^t - \boldsymbol{x}^*\|_2^2 - \frac{\mu t(t+1)}{4}\|\boldsymbol{x}^{t+1} - \boldsymbol{x}^*\|_2^2 + \frac{1}{\mu}\|\boldsymbol{g}^t\|_2^2$$

Proof of Theorem 4.4 (cont.)

Summing over all iterations before t, we get

$$\begin{aligned}
\sum_{k=0}^{t} k\left(f(\boldsymbol{x}^k) - f^{\mathrm{opt}}\right) &\leq 0 - \frac{\mu t(t+1)}{4}\|\boldsymbol{x}^{t+1} - \boldsymbol{x}^*\|_2^2 + \frac{1}{\mu}\sum_{k=0}^{t}\|\boldsymbol{g}^k\|_2^2 \\
&\leq \frac{t}{\mu}L_f^2
\end{aligned}$$

$$\Longrightarrow \quad f^{\mathsf{best},k} - f^{\mathsf{opt}} \leq \frac{L_f^2}{\mu}\frac{t}{\sum_{k=0}^{t} k} \leq \frac{2L_f^2}{\mu}\frac{1}{t+1}$$

Summary: subgradient methods

	stepsize rule	convergence rate	iteration complexity
convex & Lipschitz problems	$\eta_t \asymp \frac{1}{\sqrt{t}}$	$O\left(\frac{1}{\sqrt{t}}\right)$	$O\left(\frac{1}{\varepsilon^2}\right)$
strongly convex & Lipschitz problems	$\eta_t \asymp \frac{1}{t}$	$O\left(\frac{1}{t}\right)$	$O\left(\frac{1}{\varepsilon}\right)$

Convex-concave saddle point problems

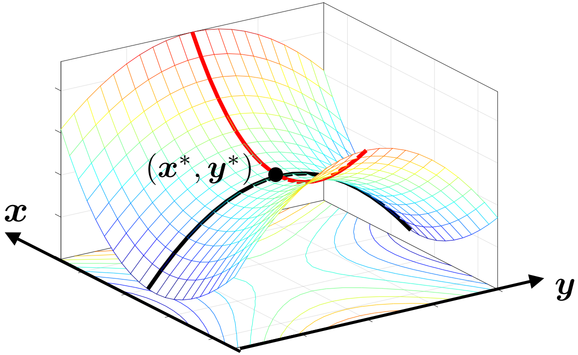
Convex-concave saddle point problems

$$\underset{\boldsymbol{x}\in\mathcal{X}}{\text{minimize}} \ \max_{\boldsymbol{y}\in\mathcal{Y}} \ f(\boldsymbol{x},\boldsymbol{y})$$

- $f(\boldsymbol{x},\boldsymbol{y})$: convex in $\boldsymbol{x}$ and concave in $\boldsymbol{y}$
- $\mathcal{X}$, $\mathcal{Y}$: bounded closed convex sets
- arises in game theory, robust optimization, generative adversarial network (GAN), ...
- under mild conditions, it is equivalent to its dual formulation

$$\underset{\boldsymbol{y}\in\mathcal{Y}}{\text{maximize}} \ \min_{\boldsymbol{x}\in\mathcal{X}} \ f(\boldsymbol{x},\boldsymbol{y})$$

Saddle points



Optimal point $(\boldsymbol{x}^*, \boldsymbol{y}^*)$ obeys

$$f(\boldsymbol{x}^*, \boldsymbol{y}) \leq f(\boldsymbol{x}^*, \boldsymbol{y}^*) \leq f(\boldsymbol{x}, \boldsymbol{y}^*), \qquad \forall \boldsymbol{x} \in \mathcal{X}, \ \boldsymbol{y} \in \mathcal{Y}$$

Projected subgradient method

A natural strategy is to apply the subgradient-based approach

$$\begin{bmatrix} \boldsymbol{x}^{t+1} \\ \boldsymbol{y}^{t+1} \end{bmatrix} = \mathcal{P}_{\mathcal{X}\times\mathcal{Y}}\left(\begin{bmatrix} \boldsymbol{x}^{t} \\ \boldsymbol{y}^{t} \end{bmatrix} - \eta_t \begin{bmatrix} \boldsymbol{g}_x^{t} \\ -\boldsymbol{g}_y^{t} \end{bmatrix} \right) \tag{4.6}$$

$$= \text{projection}\left(\begin{bmatrix} \text{subgrad descent on } \boldsymbol{x}^t \\ \text{subgrad ascent on } \boldsymbol{y}^t \end{bmatrix} \right)$$

where $\boldsymbol{g}_x^t \in \partial_{\boldsymbol{x}} f(\boldsymbol{x}^t, \boldsymbol{y}^t)$ and $-\boldsymbol{g}_y^t \in \partial_{\boldsymbol{y}} \big(- f(\boldsymbol{x}^t, \boldsymbol{y}^t) \big)$

Performance metric

One way to measure the quality of the solution is via the following error metric (think of it as a certain "duality gap")

$$\begin{aligned}\varepsilon(\boldsymbol{x},\boldsymbol{y}) &:= \left[\max_{\tilde{\boldsymbol{y}}\in\mathcal{Y}} f(\boldsymbol{x},\tilde{\boldsymbol{y}}) - f^{\text{opt}}\right] + \left[f^{\text{opt}} - \min_{\tilde{\boldsymbol{x}}\in\mathcal{X}} f(\tilde{\boldsymbol{x}},\boldsymbol{y})\right] \\ &= \max_{\tilde{\boldsymbol{y}}\in\mathcal{Y}} f(\boldsymbol{x},\tilde{\boldsymbol{y}}) - \min_{\tilde{\boldsymbol{x}}\in\mathcal{X}} f(\tilde{\boldsymbol{x}},\boldsymbol{y})\end{aligned}$$

where $f^{\text{opt}} := f(\boldsymbol{x}^*,\boldsymbol{y}^*)$ with $(\boldsymbol{x}^*,\boldsymbol{y}^*)$ the optimal solution

Convex-concave and Lipschitz problems

Theorem 4.5 (Subgradient methods for saddle point problems)

Suppose f is convex in $\boldsymbol{x}$ and concave in $\boldsymbol{y}$, and is L_f-Lipschitz continuous over $\mathcal{X} \times \mathcal{Y}$. Let $D_{\mathcal{X}}$ (resp. $D_{\mathcal{Y}}$) be the diameter of $\mathcal{X}$ (resp. $\mathcal{Y}$). Then the projected subgradient method (4.6) obeys

$$\varepsilon(\hat{\boldsymbol{x}}^t, \hat{\boldsymbol{y}}^t) \leq \frac{D_{\mathcal{X}}^2 + D_{\mathcal{Y}}^2 + L_f^2 \sum_{\tau=0}^{t} \eta_\tau^2}{2 \sum_{\tau=0}^{t} \eta_\tau}$$

where $\hat{\boldsymbol{x}}^t = \frac{\sum_{\tau=0}^{t} \eta_\tau \boldsymbol{x}^\tau}{\sum_{\tau=0}^{t} \eta_\tau}$ and $\hat{\boldsymbol{y}}^t = \frac{\sum_{\tau=0}^{t} \eta_\tau \boldsymbol{y}^\tau}{\sum_{\tau=0}^{t} \eta_\tau}$

- similar to our theory for convex problems
- suggests varying stepsize $\eta_t \asymp 1/\sqrt{t}$

Iterate averaging

Notably, it is crucial to output the weighted average $(\hat{\boldsymbol{x}}^t, \hat{\boldsymbol{y}}^t)$ of the iterates of the subgradient methods

In fact, the original iterates $(\boldsymbol{x}^t, \boldsymbol{y}^t)$ might not converge

Example (bilinear game): $f(x, y) = xy$

- When $\eta_t \to 0$ (continuous limit), (x^t, y^t) exhibits cycling behavior around $(x^*, y^*) = (0, 0)$ without converging to it

Proof of Theorem 4.5

By the convexity-concavity of f,

$$f(\boldsymbol{x}^t, \boldsymbol{y}^t) - f(\boldsymbol{x}, \boldsymbol{y}^t) \leq \langle \boldsymbol{g}_x^t, \boldsymbol{x}^t - \boldsymbol{x} \rangle, \qquad \boldsymbol{x} \in \mathcal{X}$$
$$f(\boldsymbol{x}^t, \boldsymbol{y}) - f(\boldsymbol{x}^t, \boldsymbol{y}^t) \leq \langle \boldsymbol{g}_y^t, \boldsymbol{y} - \boldsymbol{y}^t \rangle, \qquad \boldsymbol{y} \in \mathcal{Y}$$

Adding these two inequalities yields

$$f(\boldsymbol{x}^t, \boldsymbol{y}) - f(\boldsymbol{x}, \boldsymbol{y}^t) \leq \langle \boldsymbol{g}_x^t, \boldsymbol{x}^t - \boldsymbol{x} \rangle - \langle \boldsymbol{g}_y^t, \boldsymbol{y}^t - \boldsymbol{y} \rangle, \quad \boldsymbol{x} \in \mathcal{X}, \, \boldsymbol{y} \in \mathcal{Y}$$

Therefore, invoking the convexity-concavity of f once again gives

$$\begin{aligned}
\varepsilon(\hat{\boldsymbol{x}}^t, \hat{\boldsymbol{y}}^t) &= \max_{\boldsymbol{y} \in \mathcal{Y}} f(\hat{\boldsymbol{x}}^t, \boldsymbol{y}) - \min_{\boldsymbol{x} \in \mathcal{X}} f(\boldsymbol{x}, \hat{\boldsymbol{y}}^t) \\
&\leq \frac{1}{\sum_{\tau=0}^t \eta_\tau} \left\{ \max_{\boldsymbol{y} \in \mathcal{Y}} \sum_{\tau=0}^t \eta_\tau f(\boldsymbol{x}^\tau, \boldsymbol{y}) - \min_{\boldsymbol{x} \in \mathcal{X}} \sum_{\tau=0}^t \eta_\tau f(\boldsymbol{x}, \boldsymbol{y}^\tau) \right\} \\
&\leq \frac{1}{\sum_{\tau=0}^t \eta_\tau} \max_{\boldsymbol{x} \in \mathcal{X}, \boldsymbol{y} \in \mathcal{Y}} \sum_{\tau=0}^t \eta_\tau \left\{ \langle \boldsymbol{g}_x^\tau, \boldsymbol{x}^\tau - \boldsymbol{x} \rangle - \langle \boldsymbol{g}_y^\tau, \boldsymbol{y}^\tau - \boldsymbol{y} \rangle \right\} \quad (4.7)
\end{aligned}$$

Proof of Theorem 4.5 (cont.)

It then suffices to control the RHS of (4.7) as follows:

Lemma 4.6

$$\max_{\boldsymbol{x}\in\mathcal{Y},\boldsymbol{y}\in\mathcal{Y}} \sum_{\tau=0}^{t} \eta_\tau \left\{ \langle \boldsymbol{g}_x^\tau, \boldsymbol{x}^\tau - \boldsymbol{x} \rangle - \langle \boldsymbol{g}_y^\tau, \boldsymbol{y}^\tau - \boldsymbol{y} \rangle \right\}$$

$$\leq \frac{D_{\mathcal{X}}^2 + D_{\mathcal{Y}}^2 + L_f^2 \sum_{\tau=0}^{t} \eta_\tau^2}{2}$$

This lemma together with (4.7) immediately establishes Theorem 4.5

Proof of Lemma 4.6

For any $\boldsymbol{x} \in \mathcal{X}$ we have

$$\begin{aligned}
\|\boldsymbol{x}^{\tau+1} - \boldsymbol{x}\|_2^2 &= \|\mathcal{P}_{\mathcal{X}}(\boldsymbol{x}^{\tau} - \eta_\tau \boldsymbol{g}_x^{\tau}) - \mathcal{P}_{\mathcal{X}}(\boldsymbol{x})\|_2^2 \\
&\leq \|\boldsymbol{x}^{\tau} - \eta_\tau \boldsymbol{g}_x^{\tau} - \boldsymbol{x}\|_2^2 \qquad \text{(convexity of } \mathcal{X}\text{)} \\
&= \|\boldsymbol{x}^{\tau} - \boldsymbol{x}\|_2^2 - 2\eta_\tau \langle \boldsymbol{x}^{\tau} - \boldsymbol{x}, \boldsymbol{g}_x^{\tau} \rangle + \eta_\tau^2 \|\boldsymbol{g}_x^{\tau}\|_2^2
\end{aligned}$$

$$\Longrightarrow \quad 2\eta_\tau \langle \boldsymbol{x}^{\tau} - \boldsymbol{x}, \boldsymbol{g}_x^{\tau} \rangle \leq \|\boldsymbol{x}^{\tau} - \boldsymbol{x}\|_2^2 - \|\boldsymbol{x}^{\tau+1} - \boldsymbol{x}\|_2^2 + \eta_\tau^2 \|\boldsymbol{g}_x^{\tau}\|_2^2$$

Similarly, for any $\boldsymbol{y} \in \mathcal{Y}$ one has

$$-2\eta_\tau \langle \boldsymbol{y}^{\tau} - \boldsymbol{y}, \boldsymbol{g}_y^{\tau} \rangle \leq \|\boldsymbol{y}^{\tau} - \boldsymbol{y}\|_2^2 - \|\boldsymbol{y}^{\tau+1} - \boldsymbol{y}\|_2^2 + \eta_\tau^2 \|\boldsymbol{g}_y^{\tau}\|_2^2$$

Combining these two inequalities and using Lipschitz continuity yield

$$\begin{aligned}
&2\eta_\tau \langle \boldsymbol{g}_x^{\tau}, \boldsymbol{x}^{\tau} - \boldsymbol{x} \rangle - 2\eta_\tau \langle \boldsymbol{g}_y^{\tau}, \boldsymbol{y}^{\tau} - \boldsymbol{y} \rangle \\
&\quad \leq \|\boldsymbol{x}^{\tau} - \boldsymbol{x}\|_2^2 + \|\boldsymbol{y}^{\tau} - \boldsymbol{y}\|_2^2 - \|\boldsymbol{x}^{\tau+1} - \boldsymbol{x}\|_2^2 - \|\boldsymbol{y}^{\tau+1} - \boldsymbol{y}\|_2^2 + \eta_\tau^2 L_f^2
\end{aligned}$$

Proof of Lemma 4.6 (cont.)

Summing up these inequalities over $\tau = 0, \cdots, t$ gives

$$
\begin{aligned}
& 2\sum_{\tau=0}^{t} \left\{ \eta_\tau \langle \boldsymbol{g}_x^\tau, \boldsymbol{x}^\tau - \boldsymbol{x} \rangle - \eta_\tau \langle \boldsymbol{g}_y^\tau, \boldsymbol{y}^\tau - \boldsymbol{y} \rangle \right\} \\
& \leq \|\boldsymbol{x}^0 - \boldsymbol{x}\|_2^2 + \|\boldsymbol{y}^0 - \boldsymbol{y}\|_2^2 - \|\boldsymbol{x}^{t+1} - \boldsymbol{x}\|_2^2 - \|\boldsymbol{y}^{t+1} - \boldsymbol{y}\|_2^2 + L_f^2 \sum\nolimits_{\tau=0}^{t} \eta_\tau^2 \\
& \leq \|\boldsymbol{x}^0 - \boldsymbol{x}\|_2^2 + \|\boldsymbol{y}^0 - \boldsymbol{y}\|_2^2 + L_f^2 \sum\nolimits_{\tau=0}^{t} \eta_\tau^2 \\
& \leq D_{\mathcal{X}}^2 + D_{\mathcal{Y}}^2 + L_f^2 \sum\nolimits_{\tau=0}^{t} \eta_\tau^2
\end{aligned}
$$

as claimed

Remark: this lemma does NOT rely on the convexity-concavity of $f(\cdot, \cdot)$

Reference

[1] "Convex optimization, EE364B lecture notes," S. Boyd, Stanford.

[2] "Convex optimization and algorithms," D. Bertsekas, 2015.

[3] "First-order methods in optimization," A. Beck, Vol. 25, SIAM, 2017.

[4] "Convex optimization: algorithms and complexity," S. Bubeck, Foundations and trends in machine learning, 2015.

[5] "Optimization methods for large-scale systems, EE236C lecture notes," L. Vandenberghe, UCLA.

[6] "Introduction to optimization," B. Polyak, Optimization Software, 1987.

[7] "Robust stochastic approximation approach to stochastic programming," A. Nemirovski, A. Juditsky, G. Lan, A. Shapiro, SIAM Journal on optimization, 2009.