

Subgradient methods



Yuxin Chen

Princeton University, Fall 2019

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?

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

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

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

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

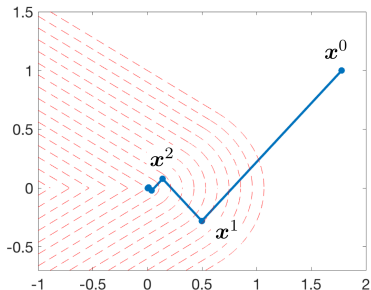
- updates

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

Issues

- Finding the steepest descent direction (or even finding a descent direction) may involve *expensive* computation
- Step size rules are tricky to choose: for certain popular step size rules (like exact line search), steepest descent might converge to non-optimal points

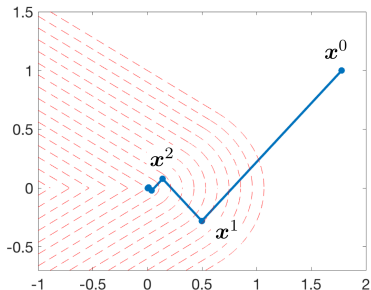
Wolfe's example



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

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

Wolfe's example



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

- 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”

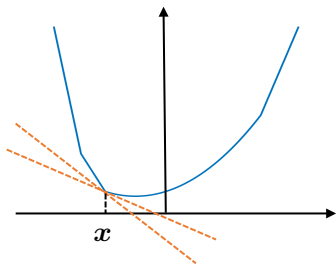
$$\mathbf{x}^{t+1} = \mathcal{P}_C(\mathbf{x}^t - \eta_t \mathbf{g}^t) \quad (4.1)$$

where \mathbf{g}^t is *any* subgradient of f at \mathbf{x}^t

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

Subgradients

Subgradients



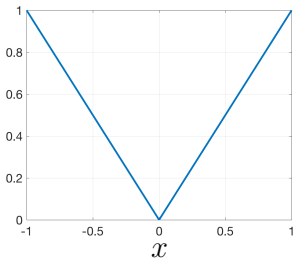
We say g is a **subgradient** of f at the point x if

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

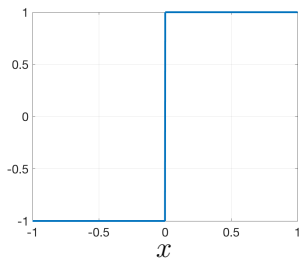
- the set of all subgradients of f at x is called the **subdifferential** of f at x , denoted by $\partial f(x)$

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

$$f(x) = |x|$$



$$\partial f(x)$$



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

Example: a subgradient of norms at 0

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

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

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

Proof: To see this, it suffices to prove that

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

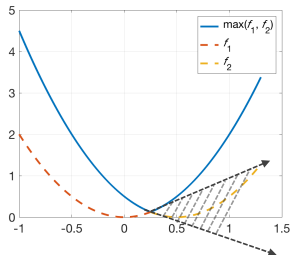
$$\iff \langle \mathbf{g}, \mathbf{z} \rangle \leq \|\mathbf{z}\|, \quad \forall \mathbf{z}$$

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

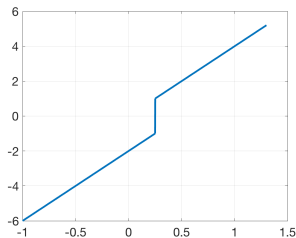
$$\langle \mathbf{g}, \mathbf{z} \rangle \leq \|\mathbf{g}\|_* \|\mathbf{z}\| \leq \|\mathbf{z}\|$$

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

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

since

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

we have

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

Basic rules (cont.)

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

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

Example: $\|\mathbf{Ax} + \mathbf{b}\|_1$

$$h(\mathbf{x}) = \|\mathbf{Ax} + \mathbf{b}\|_1$$

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

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

$$\implies \mathbf{A}^\top \mathbf{g} = \sum_{i:\mathbf{a}_i^\top \mathbf{x} + b_i \neq 0} \text{sgn}(\mathbf{a}_i^\top \mathbf{x} + b_i) \mathbf{a}_i \in \partial h(\mathbf{x})$$

Basic rules (cont.)

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

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

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

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

Basic rules (cont.)

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

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

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

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

Example: piece-wise linear functions

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

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

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

Example: the ℓ_∞ norm

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

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

$$\text{sgn}(x_j)\mathbf{e}_j \in \partial f(\mathbf{x})$$

Example: the maximum eigenvalue

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

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

Rewrite

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

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

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

Example: the nuclear norm

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

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

where $\sigma_i(\mathbf{x})$ is the i th largest singular value of \mathbf{X}

Rewrite

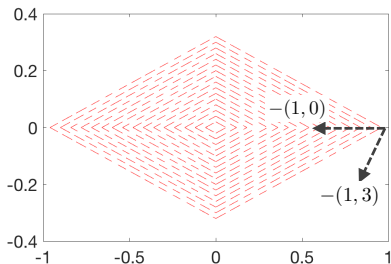
$$f(\mathbf{X}) = \sup_{\text{orthonormal } \mathbf{A}, \mathbf{B}} \langle \mathbf{A}\mathbf{B}^\top, \mathbf{X} \rangle := \sup_{\text{orthonormal } \mathbf{A}, \mathbf{B}} f_{\mathbf{A}, \mathbf{B}}(\mathbf{X})$$

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

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

Negative subgradients are not necessarily descent directions

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



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

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

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

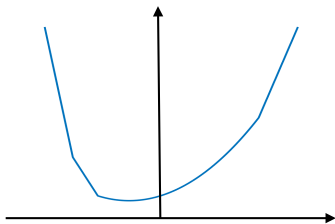
Negative subgradient is not necessarily descent direction

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

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

We also denote by $f^{\text{opt}} := \min_{\mathbf{x}} f(\mathbf{x})$ 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

$$|f(\mathbf{x}) - f(\mathbf{z})| \leq L_f \|\mathbf{x} - \mathbf{z}\|_2 \quad \forall \mathbf{x} \text{ and } \mathbf{z}$$

Fundamental inequality for projected subgradient methods

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

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

Lemma 4.1

Projected subgradient update rule (4.1) obeys

$$\|\mathbf{x}^{t+1} - \mathbf{x}^*\|_2^2 \leq \underbrace{\|\mathbf{x}^t - \mathbf{x}^*\|_2^2}_{\text{fixed}} - 2\eta_t(f(\mathbf{x}^t) - f^{\text{opt}}) + \eta_t^2 \|\mathbf{g}^t\|_2^2 \quad (4.3)$$

$\underbrace{\hspace{15em}}_{\text{majorizing function}}$

Proof of Lemma 4.1

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

where the last line uses the subgradient inequality

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

Polyak's stepsize rule

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

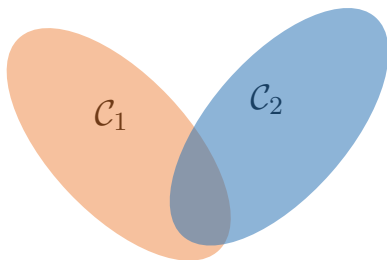
$$\eta_t = \frac{f(\mathbf{x}^t) - f^{\text{opt}}}{\|\mathbf{g}_t\|_2^2} \quad (4.4)$$

which leads to error reduction

$$\|\mathbf{x}^{t+1} - \mathbf{x}^*\|_2^2 \leq \|\mathbf{x}^t - \mathbf{x}^*\|_2^2 - \frac{(f(\mathbf{x}^t) - f(\mathbf{x}^*))^2}{\|\mathbf{g}^t\|_2^2} \quad (4.5)$$

- 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 $\mathcal{C}_1 \cap \mathcal{C}_2 \neq \emptyset$

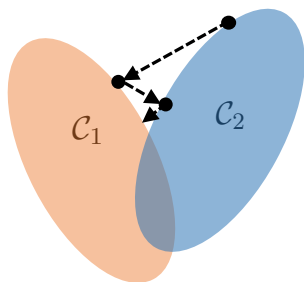
$$\text{find } \mathbf{x} \in \mathcal{C}_1 \cap \mathcal{C}_2$$



$$\text{minimize}_{\mathbf{x}} \max \{ \text{dist}_{\mathcal{C}_1}(\mathbf{x}), \text{dist}_{\mathcal{C}_2}(\mathbf{x}) \}$$

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

Example: projection onto intersection of convex sets



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

$$\mathbf{x}^{t+1} = \mathcal{P}_{C_1}(\mathbf{x}^t), \quad \mathbf{x}^{t+2} = \mathcal{P}_{C_2}(\mathbf{x}^{t+1})$$

Example: projection onto intersection of convex sets

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

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

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

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

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

which follows since $\nabla \left(\frac{1}{2} \text{dist}_{\mathcal{C}_i}^2(\mathbf{x}^t) \right) = \mathbf{x}^t - \mathcal{P}_{\mathcal{C}_i}(\mathbf{x}^t)$ (homework) and $\nabla \left(\frac{1}{2} \text{dist}_{\mathcal{C}_i}^2(\mathbf{x}^t) \right) = \text{dist}_{\mathcal{C}_i}(\mathbf{x}^t) \cdot \nabla \text{dist}_{\mathcal{C}_i}(\mathbf{x}^t)$

Example: projection onto intersection of convex sets

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

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

where $i = \arg \max_{j=1,2} \text{dist}_{C_j}(\mathbf{x}^t)$

□

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

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

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

Proof of Theorem 4.2

We have seen from (4.5) that

$$\begin{aligned}(f(\mathbf{x}^t) - f(\mathbf{x}^*))^2 &\leq \left\{ \|\mathbf{x}^t - \mathbf{x}^*\|_2^2 - \|\mathbf{x}^{t+1} - \mathbf{x}^*\|_2^2 \right\} \|\mathbf{g}^t\|_2^2 \\ &\leq \left\{ \|\mathbf{x}^t - \mathbf{x}^*\|_2^2 - \|\mathbf{x}^{t+1} - \mathbf{x}^*\|_2^2 \right\} L_f^2\end{aligned}$$

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

$$\begin{aligned}\sum_{k=0}^t (f(\mathbf{x}^k) - f(\mathbf{x}^*))^2 &\leq \left\{ \|\mathbf{x}^0 - \mathbf{x}^*\|_2^2 - \|\mathbf{x}^{t+1} - \mathbf{x}^*\|_2^2 \right\} L_f^2 \\ \implies (t+1)(f^{\text{best},t} - f^{\text{opt}})^2 &\leq \|\mathbf{x}^0 - \mathbf{x}^*\|_2^2 L_f^2\end{aligned}$$

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

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

Implications: stepsize rules

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

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

i.e. may converge to non-optimal points

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

$$\lim_{t \rightarrow \infty} f^{\text{best},t} = 0$$

i.e. converges to optimal points

Implications: stepsize rule

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

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

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

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

Rearranging terms, we are left with

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

$$\implies f^{\text{best},t} - f^{\text{opt}} \leq \frac{\sum_{i=0}^t \eta_i (f(\mathbf{x}^i) - f^{\text{opt}})}{\sum_{i=0}^t \eta_i} \leq \frac{\|\mathbf{x}^0 - \mathbf{x}^*\|_2^2 + L_f^2 \sum_{i=0}^t \eta_i^2}{2 \sum_{i=0}^t \eta_i}$$

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 diminishes 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 C . If $\eta_t \equiv \eta = \frac{2}{\mu(t+1)}$, then

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

Proof of Theorem 4.4

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

$$\|\mathbf{x}^{t+1} - \mathbf{x}^*\|_2^2 \leq (1 - \mu\eta_t)\|\mathbf{x}^t - \mathbf{x}^*\|_2^2 - 2\eta_t(f(\mathbf{x}^t) - f^{\text{opt}}) + \eta_t^2\|\mathbf{g}^t\|_2^2$$

$$\implies f(\mathbf{x}^t) - f^{\text{opt}} \leq \frac{1 - \mu\eta_t}{2\eta_t}\|\mathbf{x}^t - \mathbf{x}^*\|_2^2 - \frac{1}{2\eta_t}\|\mathbf{x}^{t+1} - \mathbf{x}^*\|_2^2 + \frac{\eta_t}{2}\|\mathbf{g}^t\|_2^2$$

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

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

and hence

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

Proof of Theorem 4.4 (cont.)

Summing over all iterations before t , we get

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

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

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

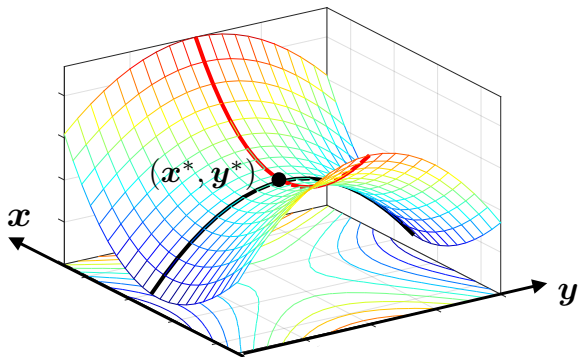
Convex-concave saddle point problems

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

- $f(\mathbf{x}, \mathbf{y})$: convex in \mathbf{x} and concave in \mathbf{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

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

Saddle points



Optimal point (x^*, y^*) obeys

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

Projected subgradient method

A natural strategy is to apply the subgradient-based approach

$$\begin{aligned} \begin{bmatrix} \mathbf{x}^{t+1} \\ \mathbf{y}^{t+1} \end{bmatrix} &= \mathcal{P}_{\mathcal{X} \times \mathcal{Y}} \left(\begin{bmatrix} \mathbf{x}^t \\ \mathbf{y}^t \end{bmatrix} - \eta_t \begin{bmatrix} \mathbf{g}_x^t \\ -\mathbf{g}_y^t \end{bmatrix} \right) \\ &= \text{projection} \left(\begin{bmatrix} \text{subgrad descent on } \mathbf{x}^t \\ \text{subgrad ascent on } \mathbf{y}^t \end{bmatrix} \right) \end{aligned} \quad (4.6)$$

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

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”)

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

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

Convex-concave and Lipschitz problems

Theorem 4.5 (Subgradient methods for saddle point problems)

Suppose f is convex in \mathbf{x} and concave in \mathbf{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

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

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

- 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 $(\hat{\mathbf{x}}^t, \hat{\mathbf{y}}^t)$ of the iterates of the subgradient methods

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

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

- When $\eta_t \rightarrow 0$ (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 ,

$$\begin{aligned} f(\mathbf{x}^t, \mathbf{y}^t) - f(\mathbf{x}, \mathbf{y}^t) &\leq \langle \mathbf{g}_x^t, \mathbf{x}^t - \mathbf{x} \rangle, & \mathbf{x} \in \mathcal{X} \\ f(\mathbf{x}^t, \mathbf{y}) - f(\mathbf{x}^t, \mathbf{y}^t) &\leq \langle \mathbf{g}_y^t, \mathbf{y} - \mathbf{y}^t \rangle, & \mathbf{y} \in \mathcal{Y} \end{aligned}$$

Adding these two inequalities yields

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

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

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

Proof of Theorem 4.5 (cont.)

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

Lemma 4.6

$$\begin{aligned} \max_{\mathbf{x} \in \mathcal{X}, \mathbf{y} \in \mathcal{Y}} \sum_{\tau=0}^t \eta_{\tau} \left\{ \langle \mathbf{g}_{\mathbf{x}}^{\tau}, \mathbf{x}^{\tau} - \mathbf{x} \rangle - \langle \mathbf{g}_{\mathbf{y}}^{\tau}, \mathbf{y}^{\tau} - \mathbf{y} \rangle \right\} \\ \leq \frac{D_{\mathcal{X}}^2 + D_{\mathcal{Y}}^2 + L_f^2 \sum_{\tau=0}^t \eta_{\tau}^2}{2} \end{aligned}$$

This lemma together with (4.7) immediately establishes Theorem 4.5

Proof of Lemma 4.6

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

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

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

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

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

Combining these two inequalities and using Lipschitz continuity yield

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

Proof of Lemma 4.6 (cont.)

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

$$\begin{aligned} & 2 \sum_{\tau=0}^t \{ \eta_{\tau} \langle \mathbf{g}_{\mathbf{x}}^{\tau}, \mathbf{x}^{\tau} - \mathbf{x} \rangle - \eta_{\tau} \langle \mathbf{g}_{\mathbf{y}}^{\tau}, \mathbf{y}^{\tau} - \mathbf{y} \rangle \} \\ & \leq \|\mathbf{x}^0 - \mathbf{x}\|_2^2 + \|\mathbf{y}^0 - \mathbf{y}\|_2^2 - \|\mathbf{x}^{t+1} - \mathbf{x}\|_2^2 - \|\mathbf{y}^{t+1} - \mathbf{y}\|_2^2 + L_f^2 \sum_{\tau=0}^t \eta_{\tau}^2 \\ & \leq \|\mathbf{x}^0 - \mathbf{x}\|_2^2 + \|\mathbf{y}^0 - \mathbf{y}\|_2^2 + L_f^2 \sum_{\tau=0}^t \eta_{\tau}^2 \\ & \leq D_{\mathcal{X}}^2 + D_{\mathcal{Y}}^2 + L_f^2 \sum_{\tau=0}^t \eta_{\tau}^2 \end{aligned}$$

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.