

CS794/CO673: Optimization for Data Science

Lec 05: Subgradient

Yaoliang Yu



UNIVERSITY OF
WATERLOO

FACULTY OF MATHEMATICS
DAVID R. CHERITON SCHOOL
OF COMPUTER SCIENCE

September 23, 2022

Nonsmooth minimization:

$$f_* = \inf_{\mathbf{w} \in C} f(\mathbf{w})$$

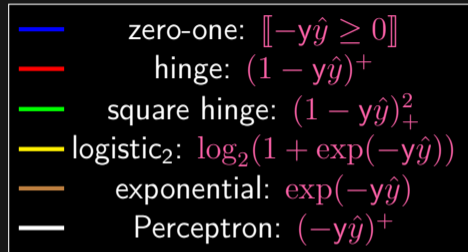
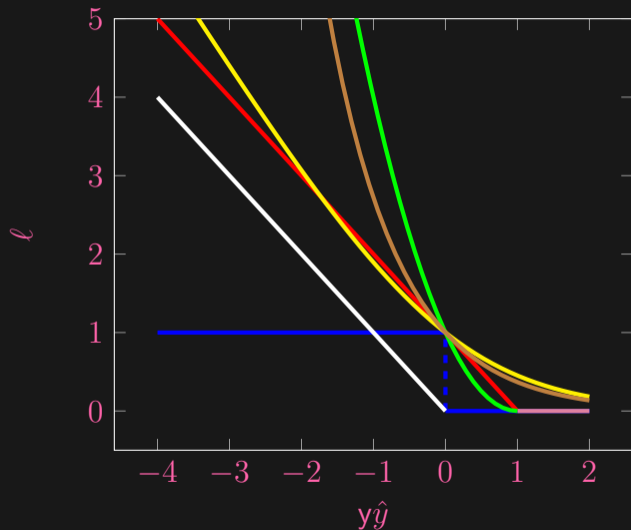
- f : nonsmooth and possibly nonconvex
- C : constraint, possibly nonconvex
- Minimizer may or may not be attained
- Maximization is just negation

Support Vector Machines

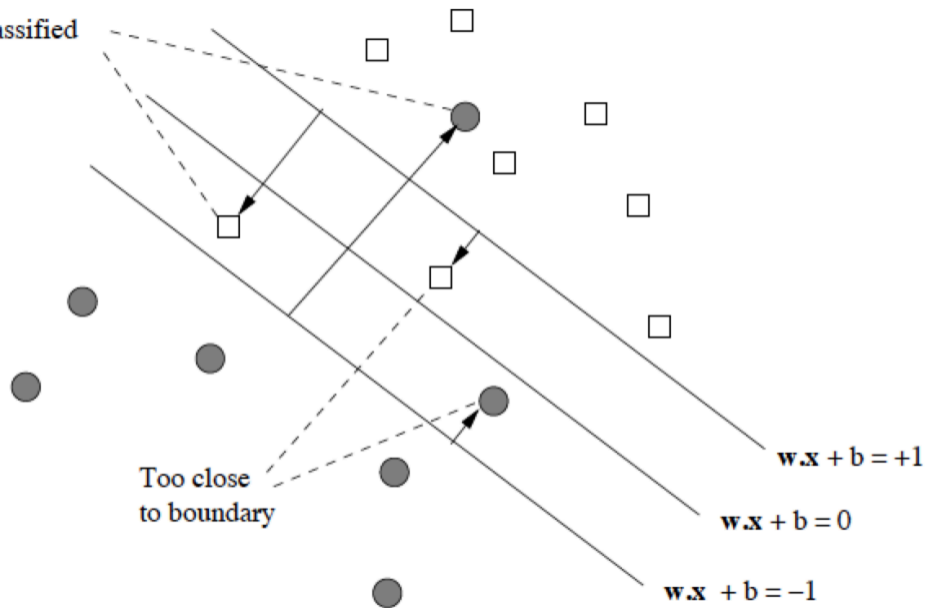
$$\min_{\mathbf{w} \in \mathbb{R}^d, b \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^n (1 - y_i \hat{y}_i)_+ + C \|\mathbf{w}\|_2^2, \quad \text{where } \hat{y}_i := \langle \mathbf{w}, \mathbf{x}_i \rangle + b,$$

- $\|\mathbf{w}\|_2^2$: margin maximization
- $(1 - y_i \hat{y}_i)_+$: i -th training error, 0 if $y_i \hat{y}_i \geq 1$ and $1 - y_i \hat{y}_i$ otherwise
- C : hyper-parameter to control tradeoff
- Cannot let $r(\mathbf{w}) = \frac{1}{n} \sum_{i=1}^n (1 - y_i \hat{y}_i)_+$ and attempt to compute P_r^η

The Hinge Loss



Misclassified



Too close
to boundary

$$w \cdot x + b = +1$$

$$w \cdot x + b = 0$$

$$w \cdot x + b = -1$$

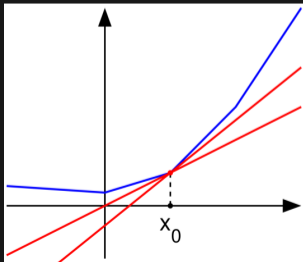
Subgradient and Subdifferential

The **subdifferential** of a **convex** function at **w** is the **set**

$$\partial f(\mathbf{w}) := \{\mathbf{g} \in \mathbb{R}^d : \forall \mathbf{z}, f(\mathbf{z}) \geq f(\mathbf{w}) + \langle \mathbf{z} - \mathbf{w}; \mathbf{g} \rangle\}$$

Any $\mathbf{g} \in \partial f(\mathbf{w})$ is called a **subgradient** of f at \mathbf{w} .

- The subdifferential is always closed and convex
- Nonempty if $\mathbf{w} \in \text{int dom } f$



Optimality Condition

Theorem: generalizing Fermat's condition

$\mathbf{w} \in \operatorname{argmin} f \implies \mathbf{0} \in \partial f(\mathbf{w})$, and the converse holds if f is convex.

- When f is continuously differentiable, then $\partial f = \nabla f$
- Necessary but not sufficient for nonconvex function
- More generally, define the “derivative” $\partial f : \mathbb{R}^d \rightarrow \mathbb{R}^d$ with some nice properties
 - reduces to the usual one if f is continuously differentiable
 - \mathbf{w} is extremal $\implies \mathbf{0} \in \partial f(\mathbf{w})$
 - nice calculus to allow practical computation

Subdifferential Calculus

Definition: Clarke's subdifferential

Locally Lipschitz continuous functions are differentiable almost everywhere, so we can define subdifferential as limits:

$$\partial f(\mathbf{w}) = \text{conv}\{\mathbf{g} : \exists \mathbf{z}_n \rightarrow \mathbf{w}, \nabla f(\mathbf{z}_n) \rightarrow \mathbf{g}\}.$$

- $\partial f(\mathbf{w}) = \nabla f(\mathbf{w})$ if f is continuously differentiable at \mathbf{w}
- $\partial(\alpha f) = \alpha \cdot \partial f$ ($\alpha > 0$ for convex functions)
- $\partial(f + g) \supseteq \partial f + \partial g$, equality holds if one of f and g is continuously differentiable
- $\partial(f \circ g) = \nabla g \cdot \partial f$ if g is continuously differentiable
- f is L -Lipschitz continuous iff $\|\partial f\| \leq L$

Example: positive part

$$\partial(t)_+ = \partial \max\{t, 0\} = \begin{cases} 1, & t > 0 \\ 0, & t < 0 \\ [0, 1], & t = 0 \end{cases}$$

Example: envelope function

Let $f(\mathbf{w}) = \max_i f_i(\mathbf{w})$ where each f_i is continuously differentiable. Then,

$$\partial f(\mathbf{w}) = \text{conv}\{\partial f_i(\mathbf{w}) : f_i(\mathbf{w}) = f(\mathbf{w})\}$$

Example: absolute function

$$\partial|t| = ?$$

The Difficulty of Nonsmoothness

- Consider the nonsmooth (separable) function

$$f(\mathbf{w}) = |w_1| + \frac{1}{2}w_2^2.$$

- The global minimizer is at $\mathbf{w} = (0, 0)$
- Let $\mathbf{w} = (0, 1)$, choose the subgradient $\mathbf{g} = (1, 1)$ and run “gradient” descent

$$\mathbf{w} \leftarrow \mathbf{w} - \eta \cdot \mathbf{g}$$

- Cauchy’s step size rule:

$$\min_{\eta \geq 0} |\eta| + \frac{1}{2}(1 - \eta)^2,$$

leading to $\eta = 0$ and we are stuck!

The Minimum Point Algorithm

Algorithm 1: The minimum-point subgradient algorithm, **may NOT converge**

Input: $\mathbf{w}_0 \in \text{dom } f$

```
1 for  $t = 0, 1, \dots$  do
2    $\mathbf{d}_t \leftarrow \underset{\mathbf{d} \in \partial f(\mathbf{w}_t)}{\text{argmin}} \|\mathbf{d}\|_2$  // choose the minimum subgradient
3   choose step size  $\eta_t$  // e.g. Cauchy's rule:  $\eta_t = \underset{\eta \geq 0}{\text{argmin}} f(\mathbf{w}_t - \eta \mathbf{d}_t)$ 
4    $\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t - \eta_t \mathbf{d}_t$ 
```

- Reduces to gradient descent if f is smooth
- Descending: $f(\mathbf{w}_{t+1}) < f(\mathbf{w}_t)$ (provided the step size is chosen suitably)
- But, it does not necessarily converge to the minimum, even under convexity!

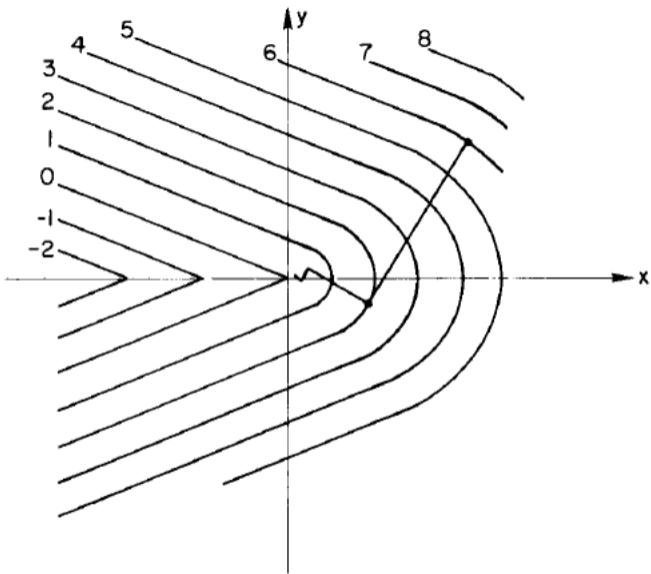


Fig. 1. Contours of f and steepest descent path.

Algorithm 2: The subgradient algorithm

Input: $\mathbf{w}_0 \in C$

```
1 for  $t = 0, 1, \dots$  do
2   choose  $\mathbf{d}_t \in \partial f(\mathbf{w}_t)$ 
3   optional:  $\mathbf{d}_t \leftarrow \mathbf{d}_t / \|\mathbf{d}_t\|_2$  // normalize
4   choose step size  $\eta_t$  // e.g.  $\eta_t = O(1/t)$ 
5    $\mathbf{w}_{t+1} \leftarrow P_C(\mathbf{w}_t - \eta_t \mathbf{d}_t)$ 
```

- $\eta_t \rightarrow 0, \sum_t \eta_t = \infty$, e.g. $\eta_t = O(1/\sqrt{t})$
- $\sum_t \eta_t = \infty, \sum_t \eta_t^2 < \infty$, e.g. $\eta_t = O(1/t)$
- $\eta_t \equiv \eta$
- $\eta_t = \eta^t$
- When the minimum value f_* is known in advance, may also use $\eta_t = \frac{f(\mathbf{w}_t) - f_*}{\|\mathbf{d}_t\|}$

To normalize or not?

Consider minimizing the convex function $f(w) = w^4$.

- With normalization: $\bar{w}_{t+1} = \bar{w}_t - \eta_t \text{sign}(\bar{w}_t) = \text{sign}(\bar{w}_t)(|\bar{w}_t| - \eta_t)$
 - $\bar{w}_t \rightarrow 0$ as long as $\eta_t \rightarrow 0$ and $\sum_t \eta_t = \infty$
- Without normalization: $w_{t+1} = w_t - 4\eta_t w_t^3 = (1 - 4\eta_t w_t^2)w_t$
 - if we start with $w_1 = 1$ and $\eta_t = 1/t$, then
$$w_t^2 \geq 1/\eta_t \implies w_{t+1}^2 = (4\eta_t w_t^2 - 1)^2 w_t^2 \geq (4w_t - 1)^2 w_t^2 \geq 9w_t^2 \geq 9t \geq t + 1 = 1/\eta_{t+1},$$

i.e. $|w_t| \rightarrow \infty$.

Nonexpansion

A mapping $T : \mathbb{R}^d \rightarrow \mathbb{R}^d$ is called a nonexpansion iff it is 1-Lipschitz continuous:

$$\|T\mathbf{w} - T\mathbf{z}\| \leq \|\mathbf{w} - \mathbf{z}\|$$

Almost all algorithms in this course can be written abstractly as

$$\mathbf{w}_{t+1} \leftarrow T_t \mathbf{w}_t,$$

where the mapping T_t often is a nonexpansion (and may not depend on t).

Theorem: Euclidean projection to convex sets is nonexpansion

Let C be a (closed) convex set. Then P_C is nonexpansive:

$$\|P_C(\mathbf{w}) - P_C(\mathbf{z})\|_2 \leq \|\mathbf{w} - \mathbf{z}\|_2.$$

Same is true for the proximal map P_f^η when f is convex.

Theorem: convergence of subgradient

Let $C \subseteq \mathbb{R}^d$ be (closed) convex and $f : C \rightarrow \mathbb{R}$ be L -Lipschitz continuous convex (w.r.t. $\|\cdot\|_2$). For any $\mathbf{w} \in C$, subgradient (without normalization) satisfies:

$$\min_{0 \leq t \leq T-1} f(\mathbf{w}_t) - f(\mathbf{w}) \leq \frac{\sum_{t=0}^{T-1} \eta_t}{\sum_{s=0}^{T-1} \eta_s} (f(\mathbf{w}_t) - f(\mathbf{w})) \leq \frac{\|\mathbf{w}_0 - \mathbf{w}\|_2^2 + L^2 \sum_{t=0}^{T-1} \eta_t^2}{2 \sum_{s=0}^{T-1} \eta_s}.$$

- RHS vanishes iff $\sum_{s=0}^{T-1} \eta_s = \infty$ and $\sum_{t=0}^{T-1} \eta_t^2 < \infty$ iff $\eta_t \rightarrow 0$, $\sum_{s=0}^{T-1} \eta_s = \infty$.
- Fix accuracy ϵ , can set $\eta_t = \eta = \frac{\epsilon}{L^2}$ and obtain $T = \frac{L^2 \|\mathbf{w}_0 - \mathbf{w}\|_2^2}{\epsilon^2}$ iterations suffice
- No explicit dependence on dimension d
- Slower than $O(\frac{1}{\epsilon})$ of gradient descent: price of nonsmoothness

$$\begin{aligned}
\|\mathbf{w}_{t+1} - \mathbf{w}\|_2^2 &= \|\mathbf{P}_C(\mathbf{w}_t - \eta_t \mathbf{d}_t) - \mathbf{w}\|_2^2 \\
[\mathbf{w} \in C] &= \|\mathbf{P}_C(\mathbf{w}_t - \eta_t \mathbf{d}_t) - \mathbf{P}_C(\mathbf{w})\|_2^2 \\
[\text{projections are nonexpansive}] &\leq \|\mathbf{w}_t - \eta_t \mathbf{d}_t - \mathbf{w}\|_2^2 \\
&= \|\mathbf{w}_t - \mathbf{w}\|_2^2 + \eta_t^2 \|\mathbf{d}_t\|_2^2 - 2\eta_t \langle \mathbf{w}_t - \mathbf{w}, \mathbf{d}_t \rangle \\
[\mathbf{d}_t \text{ is a subgradient, } \eta_t \geq 0] &\leq \|\mathbf{w}_t - \mathbf{w}\|_2^2 + \eta_t^2 \|\mathbf{d}_t\|_2^2 + 2\eta_t (f(\mathbf{w}) - f(\mathbf{w}_t)) \\
[\partial f \text{ is bounded by } L] &\leq \|\mathbf{w}_t - \mathbf{w}\|_2^2 + \eta_t^2 L^2 + 2\eta_t (f(\mathbf{w}) - f(\mathbf{w}_t)).
\end{aligned}$$

Telescoping we obtain:

$$\begin{aligned}
\|\mathbf{w}_T - \mathbf{w}\|_2^2 &\leq \|\mathbf{w}_0 - \mathbf{w}\|_2^2 + L^2 \sum_{t=0}^{T-1} \eta_t^2 + 2 \sum_{t=0}^{T-1} \frac{\eta_t}{\sum_{s=0}^{T-1} \eta_s} (f(\mathbf{w}) - f(\mathbf{w}_t)) \cdot \sum_{s=0}^{T-1} \eta_s \\
\min_{0 \leq t \leq T-1} f(\mathbf{w}_t) - f(\mathbf{w}) &\leq \sum_{t=0}^{T-1} \frac{\eta_t}{\sum_{s=0}^{T-1} \eta_s} (f(\mathbf{w}_t) - f(\mathbf{w})) \leq \frac{\|\mathbf{w}_0 - \mathbf{w}\|_2^2 + L^2 \sum_{t=0}^{T-1} \eta_t^2}{2 \sum_{s=0}^{T-1} \eta_s}
\end{aligned}$$

Extending to Composite

$$\min_{\mathbf{w}} f(\mathbf{w}), \quad \text{where } f(\mathbf{w}) = \ell(\mathbf{w}) + r(\mathbf{w})$$

Algorithm 3: The proximal subgradient algorithm

Input: \mathbf{w}_0 , functions ℓ and r

```
1 for  $t = 0, 1, \dots$  do
2   choose  $\mathbf{d}_t \in \partial\ell(\mathbf{w}_t)$ 
3   optional:  $\mathbf{d}_t \leftarrow \mathbf{d}_t / \|\mathbf{d}_t\|_2$  // normalize
4   choose step size  $\eta_t$  // e.g.  $\eta_t = O(1/t)$ 
5    $\mathbf{z}_{t+1} \leftarrow \mathbf{w}_t - \eta_t \mathbf{d}_t$  // subgradient w.r.t.  $\ell$ 
6    $\mathbf{w}_{t+1} \leftarrow P_r^{\eta_t}(\mathbf{z}_{t+1})$  // proximal w.r.t.  $r$ 
```

Example: Elastic net

$$\min_{\mathbf{w}} \frac{1}{n} \|\mathbf{w}\mathbf{X} - \mathbf{y}\|_2^2 + \lambda \|\mathbf{w}\|_1 + \frac{\gamma}{2} \|\mathbf{w}\|_2^2$$

Now we have 4 choices:

- Set $\ell = \frac{1}{n} \|\mathbf{w}\mathbf{X} - \mathbf{y}\|_2^2 + \frac{\gamma}{2} \|\mathbf{w}\|_2^2$ and $r = \lambda \|\mathbf{w}\|_1$.
- Set $\ell = \frac{1}{n} \|\mathbf{w}\mathbf{X} - \mathbf{y}\|_2^2$ and $r = \lambda \|\mathbf{w}\|_1 + \frac{\gamma}{2} \|\mathbf{w}\|_2^2$.
- Set $\ell = \frac{1}{n} \|\mathbf{w}\mathbf{X} - \mathbf{y}\|_2^2 + \lambda \|\mathbf{w}\|_1 + \frac{\gamma}{2} \|\mathbf{w}\|_2^2$.
- Set $\ell = \frac{1}{n} \|\mathbf{w}\mathbf{X} - \mathbf{y}\|_2^2 + \lambda \|\mathbf{w}\|_1$ and $r = \frac{\gamma}{2} \|\mathbf{w}\|_2^2$.

What are the pros and cons?

