Optimization for Data Science

Lec 04: Subgradient

Yaoliang Yu

UNIVERSITY OF WATERLOO | FACULTY OF MATHEMATICS
DAVID R. CHERITON SCHOOL OF COMPUTER SCIENCE

Problem

Nonsmooth minimization:

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

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

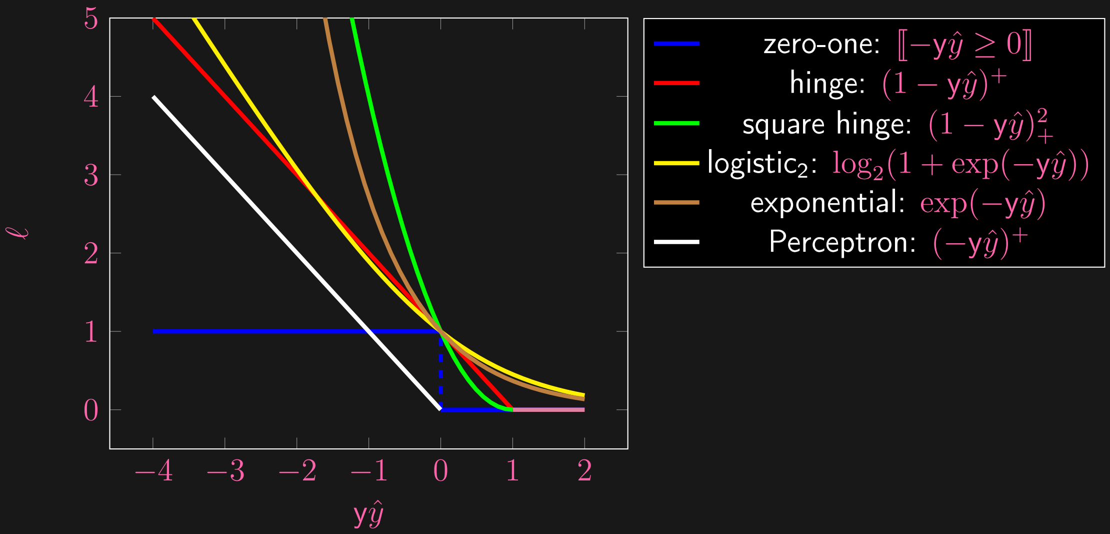
Support Vector Machines

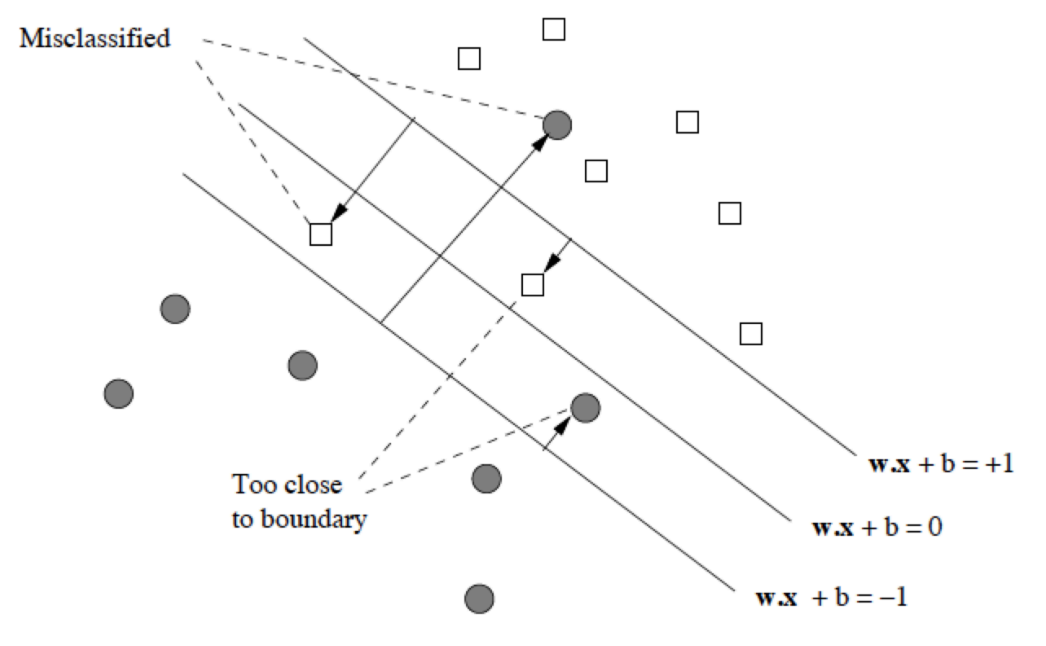
$$\min_{\mathbf{w}\in\mathbb{R}^d, b\in\mathbb{R}} \ \frac{1}{n}\sum_{i=1}^{n}(1 - y_i\hat{y}_i)_+ + C\|\mathbf{w}\|_2^2, \quad \text{where} \quad \hat{y}_i := \langle \mathbf{w}, \mathbf{x}_i \rangle + b,$$

- $\|\mathbf{w}\|_2^2$: margin maximization
- $(1 - \mathrm{y}_i\hat{y}_i)^+$: i-th training error, 0 if $\mathrm{y}_i\hat{y}_i \geq 1$ and $1 - \mathrm{y}_i\hat{y}_i$ 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

C. Cortes and V. Vapnik. "Support-vector networks". *Machine Learning*, vol. 20, no. 3 (1995), pp. 273–297.

The Hinge Loss





Misclassified
Too close
to boundary
w.x + b = +1
w.x + b = 0
w.x + b = −1

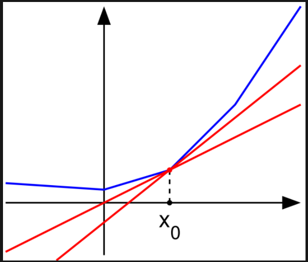
Subgradient and Subdifferential

The subdifferential of a convex function at $\mathbf{w}$ 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 $\mathbf{w} \in \operatorname{int} \operatorname{dom} f$



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" $\partial f : \mathbb{R}^d \to \mathbb{R}^d$ 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:

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

- $\partial f(\mathbf{w}) = \nabla f(\mathbf{w})$ 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 $\|\partial f\| \leq \mathsf{L}$

F. H. Clarke. "Optimization and Nonsmooth Analysis". reprinted from the 1983 edition. SIAM, 1990.

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,

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

Example: absolute function

$$\partial |t| = ?$$

The Difficulty of Nonsmoothness

- Consider the nonsmooth (separable) function

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

- 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:

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

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

The Minimum Point Algorithm

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

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

```
for t = 0, 1, ... do
    d_t ← argmin_{d ∈ ∂f(w_t)} ||d||_2          // choose the minimum subgradient
    choose step size η_t                          // e.g. Cauchy's rule: η_t = argmin_{η≥0} f(w_t − η d_t)
    w_{t+1} ← w_t − η_t d_t
```

$$\mathbf{d}_t \leftarrow \operatorname*{argmin}_{\mathbf{d}\in\partial f(\mathbf{w}_t)} \|\mathbf{d}\|_2$$

$$\eta_t = \operatorname*{argmin}_{\eta\geq 0} f(\mathbf{w}_t - \eta_t \mathbf{d}_t)$$

$$\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t - \eta_t \mathbf{d}_t$$

- 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!

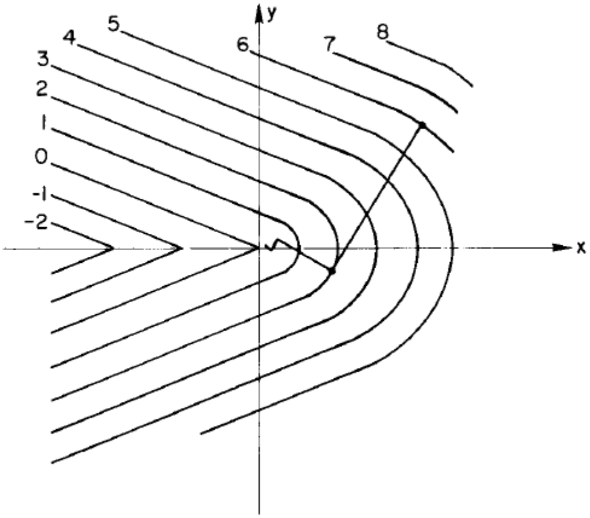


Fig. 1. Contours of f and steepest descent path.

P. Wolfe. "A method of conjugate subgradients for minimizing nondifferentiable functions". *Mathematical Programming Study*, vol. 3 (1975), pp. 145–173.

Algorithm 2: The subgradient algorithm

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

for $t = 0, 1, \ldots$ **do**

  choose $\mathbf{d}_t \in \partial f(\mathbf{w}_t)$

  optional: $\mathbf{d}_t \leftarrow \mathbf{d}_t / \|\mathbf{d}_t\|_2$ // normalize

  choose step size η_t // e.g. $\eta_t = O(1/t)$

  $\mathbf{w}_{t+1} \leftarrow \mathrm{P}_C(\mathbf{w}_t - \eta_t \mathbf{d}_t)$

- $\eta_t \to 0, \sum_t \eta_t = \infty$, 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 $f_\star$ is known in advance, may also use $\eta_t = \frac{f(\mathbf{w}_t) - f_\star}{\|\mathbf{d}_t\|}$

B. Polyak. "Minimization of unsmooth functionals". *USSR Computational Mathematics and Mathematical Physics*, vol. 9, no. 3 (1969), pp. 14–29.

To normalize or not?

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

- With normalization: $\bar{w}_{t+1} = \bar{w}_t - \eta_t \operatorname{sign}(\bar{w}_t) = \operatorname{sign}(\bar{w}_t)(|\bar{w}_t| - \eta_t)$
  - $\bar{w}_t \to 0$ as long as $\eta_t \to 0$ 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. $|w_t| \to \infty$.

Nonexpansion

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

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

Almost all algorithms in this course can be written abstractly as

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

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:

$$\|\mathrm{P}_C(\mathbf{w}) - \mathrm{P}_C(\mathbf{z})\|_2 \leq \|\mathbf{w} - \mathbf{z}\|_2.$$

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

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

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

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

Telescoping we obtain:

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

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

Extending to Composite

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

Algorithm 3: The proximal subgradient algorithm

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

```
for t = 0, 1, ... do
    choose d_t ∈ ∂ℓ(w_t)
    optional: d_t ← d_t/‖d_t‖_2                    // normalize
    choose step size η_t                           // e.g. η_t = O(1/t)
    z_{t+1} ← w_t − η_t d_t                        // subgradient w.r.t. ℓ
    w_{t+1} ← P_r^{η_t}(z_{t+1})                   // proximal w.r.t. r
```

J. C. Duchi and Y. Singer. "Efficient Online and Batch Learning Using Forward Backward Splitting". *Journal of Machine Learning Research*, vol. 10 (2009), pp. 2899–2934.

Example: Elastic net

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

Now we have 4 choices:

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

What are the pros and cons?

H. Zou and T. Hastie. "Regularization and variable selection via the elastic net". *Journal of the Royal Statistical Society, Series B*, vol. 67 (2005), pp. 301–320.