Optimization for Data Science

Lec 01: Gradient Descent

Yaoliang Yu

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

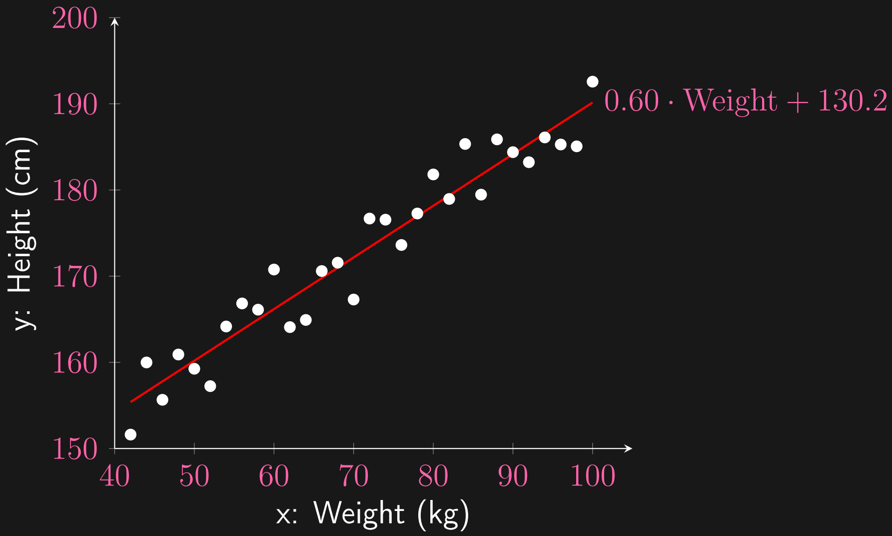
Problem

Unconstrained smooth minimization:

$$f_\star = \inf_{\mathbf{w}\in\mathbb{R}^d} f(\mathbf{w}).$$

- No constraint on the domain
- $f : \mathbb{R}^d \to \mathbb{R}$ is smooth, e.g. continuously differentiable
- f can be convex or nonconvex
- Minimizer may or may not be attained
- Maximization is just negation

Linear Regression



Linear Least Squares Regression

$$\min_{\mathbf{w}} \frac{1}{n} \sum_{i=1}^{n} (\langle \mathbf{x}_i, \mathbf{w} \rangle - y_i)^2 \quad \equiv \quad \min_{\mathbf{w}} \underbrace{\tfrac{1}{n} \|\mathbf{w}\mathbf{X} - \mathbf{y}\|_2^2}_{f(\mathbf{w})}$$

- $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \mathbb{R}^{p \times n}$
- $\mathbf{y} = [y_1, \ldots, y_n] \in \mathbb{R}^n$
- $\mathbf{w} \in \mathbb{R}^p$
- Clearly, f is quadratic and hence (continuously) differentiable
- No constraint on $\mathbf{w}$

Logistic Regression

$$\inf_{\mathbf{w}} \frac{1}{n}\sum_{i=1}^{n} \log[1+\exp(-y_i \langle \mathbf{x}_i, \mathbf{w}\rangle)] \equiv \inf_{\mathbf{w}} \underbrace{\left\langle \log[1+\exp(-\mathbf{w}\mathbf{A})], \tfrac{1}{n}\cdot \mathbf{1}\right\rangle}_{f(\mathbf{w})}$$

- $\mathbf{A} = [y_1\mathbf{x}_1, \ldots, y_n\mathbf{x}_n] \in \mathbb{R}^{p\times n}$
- $\mathbf{y} = [y_1, \ldots, y_n] \in \{\pm 1\}^n$
- $\mathbf{w} \in \mathbb{R}^p$
- Again, f is (continuously) differentiable
- No constraint on $\mathbf{w}$

Calculus Detour

(Fréchet) Derivative f' of a function f at $\mathbf{w}$:

$$\lim_{\mathbf{0} \neq \mathbf{z} \to \mathbf{0}} \frac{\|f(\mathbf{w}+\mathbf{z}) - f(\mathbf{w}) - f'(\mathbf{w})(\mathbf{z})\|}{\|\mathbf{z}\|} \to 0$$

- $f : \mathcal{X} \to \mathcal{Y} \implies f'(\mathbf{w}) : \mathcal{X} \to \mathcal{Y} \implies f' : \mathcal{X} \to (\mathcal{X} \to \mathcal{Y})$
- $f'(\mathbf{w})(\mathbf{z})$ is linear in $\mathbf{z}$ but possibly nonlinear in $\mathbf{w}$

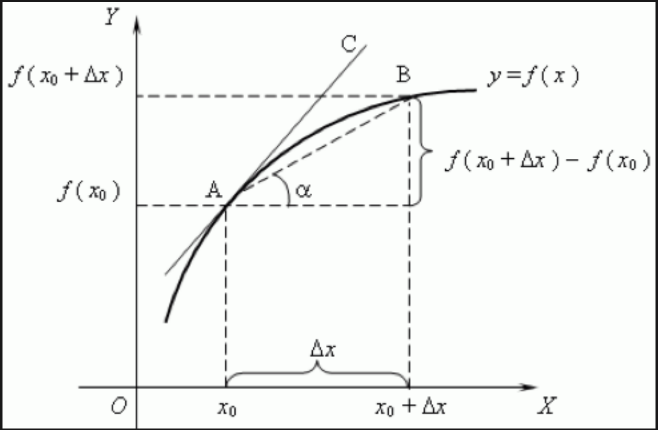
Example: Quadratic function $f(\mathbf{w}) = \langle \mathbf{w}, A\mathbf{w} + \mathbf{b} \rangle + c$

$$f(\mathbf{w}+\mathbf{z}) = \langle \mathbf{w}+\mathbf{z}, A\mathbf{w} + A\mathbf{z} + \mathbf{b} \rangle + c$$

$$f(\mathbf{w}+\mathbf{z}) - f(\mathbf{w}) = \langle \mathbf{w}, A\mathbf{z} \rangle + \langle \mathbf{z}, A\mathbf{w} \rangle + \langle \mathbf{z}, A\mathbf{z} \rangle + \langle \mathbf{z}, \mathbf{b} \rangle$$

$$f'(\mathbf{w})(\mathbf{z}) = \left\langle (A + A^\top)\mathbf{w} + \mathbf{b}, \mathbf{z} \right\rangle$$

$$f'(\mathbf{w}) = (A + A^\top)\mathbf{w} + \mathbf{b}$$



- Chain rule: $(f \circ g)'(\mathbf{w})(\mathbf{z}) = f'[g(\mathbf{w})][g'(\mathbf{w})(\mathbf{z})]$
- Often suffices to take: $[f'(\mathbf{w})]_j = \partial_j f(w_1, \ldots, w_j, \ldots, w_d)$

Example: Logistic Loss

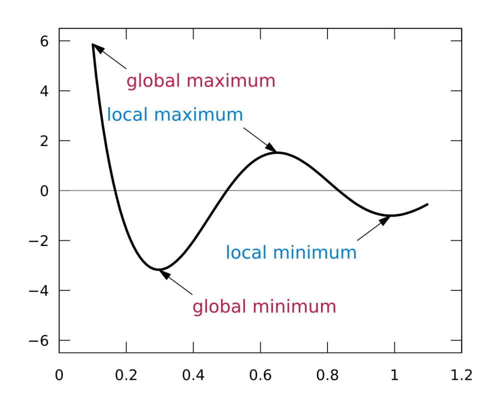
$$
\begin{aligned}
f(\mathbf{w}) &= \left\langle \log[1+\exp(-\mathbf{w}\mathbf{A})], \tfrac{1}{n}\cdot\mathbf{1}\right\rangle \\
\mathrm{d}f(\mathbf{w}) &= \left\langle \mathrm{d}\log[1+\exp(-\mathbf{w}\mathbf{A})], \tfrac{1}{n}\cdot\mathbf{1}\right\rangle + \left\langle \log[1+\exp(-\mathbf{w}\mathbf{A})], \mathrm{d}\tfrac{1}{n}\cdot\mathbf{1}\right\rangle \\
&= \left\langle \frac{-\exp(-\mathbf{w}\mathbf{A})}{1+\exp(-\mathbf{w}\mathbf{A})}\mathrm{d}\mathbf{w}\cdot\mathbf{A}, \tfrac{1}{n}\cdot\mathbf{1}\right\rangle \\
&= \left\langle \mathrm{d}\mathbf{w}, \frac{-\exp(-\mathbf{w}\mathbf{A})}{1+\exp(-\mathbf{w}\mathbf{A})}\cdot\tfrac{1}{n}\cdot\mathbf{1}\mathbf{A}^\top\right\rangle \\
\nabla f(\mathbf{w}) &= \frac{\mathrm{d}f(\mathbf{w})}{\mathrm{d}\mathbf{w}} = \frac{1}{n}\cdot\frac{-\exp(-\mathbf{w}\mathbf{A})}{1+\exp(-\mathbf{w}\mathbf{A})}\mathbf{A}^\top
\end{aligned}
$$

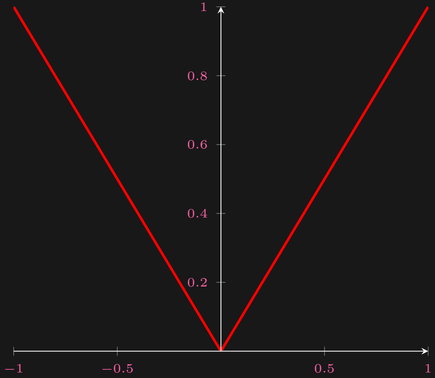
- Recall $\mathbf{w}\in\mathbb{R}^p$, $\mathbf{A}\in\mathbb{R}^{p\times n}$
- What is the dimension of our gradient $\nabla f(\mathbf{w})$?

Optimality Condition

Theorem: Fermat's necessary condition for extremity

If $\mathbf{w}$ is a minimizer (or maximizer) of a differentiable function f over an open set, then $f'(\mathbf{w}) = \mathbf{0}$.





Gradient Descent

Algorithm 1: Richardson's first-order extrapolation for linear systems

Input: $\mathbf{w}_0 \in \mathbb{R}^d$, $A \in \mathbb{R}^{d \times d}$, $\mathbf{b} \in \mathbb{R}^d$

```
for t = 0, 1, ... do
    g_t ← A w_t − b              // ‘‘gradient’’
    w_{t+1} ← w_t − η_t g_t      // η_t is the step size
```

- **for** $t = 0, 1, \ldots$ **do**
  - $\mathbf{g}_t \leftarrow A\mathbf{w}_t - \mathbf{b}$ // ‘‘gradient’’
  - $\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t - \eta_t \mathbf{g}_t$ // η_t is the step size

Algorithm 2: Gradient descent for unconstrained smooth minimization

Input: $\mathbf{w}_0 \in \mathbb{R}^d$, smooth function $f : \mathbb{R}^d \to \mathbb{R}$

- **for** $t = 0, 1, \ldots$ **do**
  - $\mathbf{g}_t \leftarrow \nabla f(\mathbf{w}_t)$ // compute the gradient
  - $\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t - \eta_t \mathbf{g}_t$ // η_t is the step size

- Repeatedly subtract a multiple of the gradient

Intuition

$$
\begin{aligned}
f(\mathbf{w}_{t+1}) &= f\big(\mathbf{w}_t - \eta_t \nabla f(\mathbf{w}_t)\big) \\
&= f(\mathbf{w}_t) - \eta_t \langle \nabla f(\mathbf{w}_t), \nabla f(\mathbf{w}_t) \rangle + o(\eta_t) \\
&= f(\mathbf{w}_t) - \eta_t \underbrace{\|\nabla f(\mathbf{w}_t)\|_2^2}_{\geq 0} + o(\eta_t)
\end{aligned}
$$

- If $\nabla f(\mathbf{w}_t) = 0$, we are done
- Otherwise for small $\eta_t > 0$, we have $f(\mathbf{w}_{t+1}) < f(\mathbf{w}_t)$
- Strict improvement at each iteration; is it enough??

Lipschitz Continuity = Bounded Derivative

Theorem:

Let $\mathsf{T} : \mathbb{R}^d \to \mathbb{R}^m$ be differentiable. Then, T is L-Lipschitz continuous:

$$\|\mathsf{T}(\mathbf{w}) - \mathsf{T}(\mathbf{z})\| \le \mathsf{L}\|\mathbf{w} - \mathbf{z}\|$$

if and only if

$$\sup_{\mathbf{w}} \|\mathsf{T}'(\mathbf{w})\| = \sup_{\mathbf{w}} \sup_{\|\mathbf{z}\| \le 1} \|\mathsf{T}'(\mathbf{w})(\mathbf{z})\| \le \mathsf{L}.$$

- Lipschitz continuity: output change is bounded by input change
- Equivalently, derivative (i.e. infinitesimal change) is bounded

L-smoothness

We call a differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ L-smooth if for all $\mathbf{w}$ and $\mathbf{z}$:

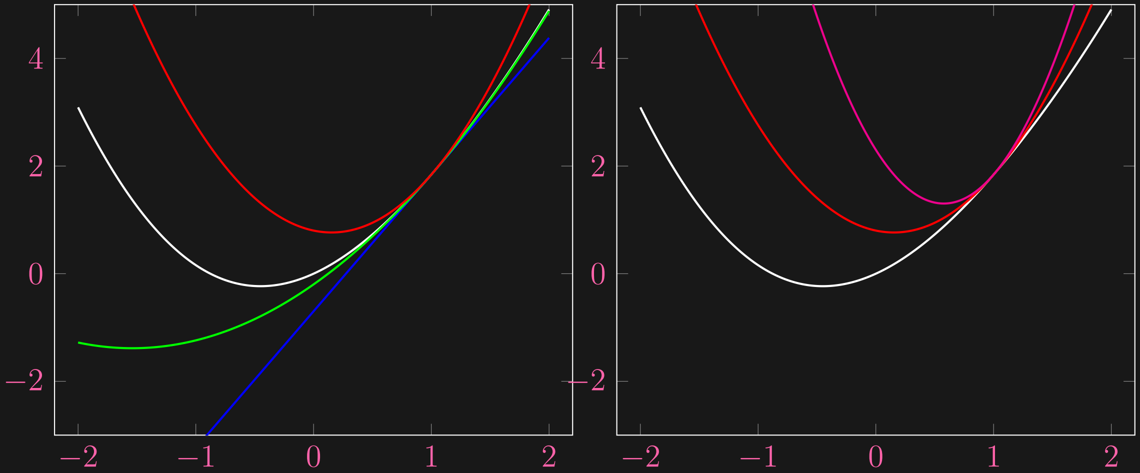
$$f(\mathbf{z}) \leq f(\mathbf{w}) + \underbrace{f'(\mathbf{w})(\mathbf{z}-\mathbf{w})}_{\langle \mathbf{z}-\mathbf{w}, \nabla f(\mathbf{w}) \rangle} + \frac{\mathsf{L}}{2}\|\mathbf{z}-\mathbf{w}\|^2$$

Theorem: Characterizing L-smoothness

Consider the following statements for a real-valued smooth function:

(I). Vector-valued derivative $f' : \mathbb{R}^d \to \mathbb{R}^d$ is L-Lipschitz continuous
(II). Matrix-valued second-order derivative $f'' : \mathbb{R}^d \to \mathbb{R}^{d \times d}$ is L-bounded
(III). Real-valued functions $\pm f$ are L-smooth

Then, (I) $\iff$ (II) $\implies$ (III). If f is convex or the norm is Euclidean, then all three are equivalent.


4
2
0
−2
−2
−1
0
1
2
4
2
0
−2
−2
−1
0
1
2

Importance of L-smoothness

$$f(\mathbf{w}) \leq f(\mathbf{w}_t) + \langle \mathbf{w} - \mathbf{w}_t, \nabla f(\mathbf{w}_t) \rangle + \frac{1}{2\eta_t} \|\mathbf{w} - \mathbf{w}_t\|_2^2$$

- RHS is a quadratic function of $\mathbf{w}$
- Equality holds if $\eta_t \leq \frac{1}{\mathsf{L}}$
- Minimize RHS w.r.t. $\mathbf{w}$:

$$\mathbf{w}_{t+1} \leftarrow \underset{\mathbf{w}}{\operatorname{argmin}}\, f(\mathbf{w}_t) + \frac{1}{2\eta_t} \|\mathbf{w} - [\mathbf{w}_t - \eta_t \nabla f(\mathbf{w}_t)]\|_2^2 - \frac{\eta_t}{2} \|\nabla f(\mathbf{w}_t)\|_2^2$$

- This is exactly gradient descent
- Moreover, $f(\mathbf{w}_{t+1}) \leq f(\mathbf{w}_t) - \frac{\eta_t}{2} \|\nabla f(\mathbf{w}_t)\|_2^2$

Example: Logistic Loss

$$f(\mathbf{w}) = \left\langle \log[1+\exp(-\mathbf{w}\mathbf{A})], \tfrac{1}{n}\cdot\mathbf{1}\right\rangle$$

$$\nabla f(\mathbf{w}) = \frac{\mathrm{d}f(\mathbf{w})}{\mathrm{d}\mathbf{w}} = \frac{1}{n}\cdot\frac{-\exp(-\mathbf{w}\mathbf{A})}{1+\exp(-\mathbf{w}\mathbf{A})}\mathbf{A}^\top = \frac{1}{n}\cdot\left[\frac{1}{1+\exp(-\mathbf{w}\mathbf{A})}-1\right]\mathbf{A}^\top$$

$$\mathrm{d}\nabla f(\mathbf{w}) = \frac{1}{n}\,\mathrm{d}\frac{1}{1+\exp(-\mathbf{w}\mathbf{A})}\cdot\mathbf{A}^\top = \frac{1}{n}\frac{\exp(-\mathbf{w}\mathbf{A})}{[1+\exp(-\mathbf{w}\mathbf{A})]^2}\,\mathrm{d}\mathbf{w}\mathbf{A}\cdot\mathbf{A}^\top$$

$$= \mathrm{d}\mathbf{w}\cdot\frac{1}{n}\mathbf{A}\,\mathrm{diag}\left(\frac{\exp(-\mathbf{w}\mathbf{A})}{[1+\exp(-\mathbf{w}\mathbf{A})]^2}\right)\mathbf{A}^\top$$

$$\nabla^2 f(\mathbf{w}) = \frac{1}{n}\mathbf{A}\,\mathrm{diag}\left(\frac{\exp(-\mathbf{w}\mathbf{A})}{[1+\exp(-\mathbf{w}\mathbf{A})]^2}\right)\mathbf{A}^\top \preceq \frac{1}{n}\mathbf{A}\mathbf{A}^\top$$

$$\sup_{\mathbf{w}}\|\nabla^2 f(\mathbf{w})\|_{\mathrm{sp}} \le \|\tfrac{1}{n}\mathbf{A}\mathbf{A}^\top\|_{\mathrm{sp}} = \tfrac{1}{n}\|\mathbf{A}\|_{\mathrm{sp}}^2 \le \tfrac{1}{n}\|\mathbf{A}\|_{\mathsf{F}}^2$$

Theorem: Convergence of gradient descent for L-smooth functions

Let $f : \mathbb{R}^d \to \mathbb{R}$ be L-smooth and bounded from below (i.e. $f_\star > -\infty$). If the step size $\eta_t \in [\alpha, \frac{2}{\mathsf{L}} - \beta]$ for some $\alpha, \beta > 0$, then the gradient descent iterate $\{\mathbf{w}_t\}$ satisfies $\nabla f(\mathbf{w}_t) \to \mathbf{0}$. Moreover,

$$\min_{0 \le t \le T-1} \|\nabla f(\mathbf{w}_t)\|_2 \le \sqrt{\frac{2[f(\mathbf{w}_0) - f_\star]}{\alpha\beta\mathsf{L}T}}.$$

Can tune α and β to optimize the bound: since $\alpha + \beta \le \frac{2}{\mathsf{L}}$, the minimum is achieved when $\alpha = \beta = \frac{1}{\mathsf{L}}$, and the bound reduces to

$$\min_{0 \le t \le T-1} \|\nabla f(\mathbf{w}_t)\|_2 \le \sqrt{\frac{2\mathsf{L}[f(\mathbf{w}_0) - f_\star]}{T}},$$

B. T. Polyak. "Gradient methods for the minimization of functionals". *USSR Computational Mathematics and Mathematical Physics*, vol. 3, no. 4 (1963), pp. 643–653.

Proof

$$f(\mathbf{w}_{t+1}) = f(\mathbf{w}_t - \eta_t \nabla f(\mathbf{w}_t)) \le f(\mathbf{w}_t) - \eta_t \|\nabla f(\mathbf{w}_t)\|_2^2 + \frac{\mathsf{L}\eta_t^2}{2} \|\nabla f(\mathbf{w}_t)\|_2^2$$
$$= f(\mathbf{w}_t) - \eta_t (1 - \frac{\mathsf{L}\eta_t}{2}) \|\nabla f(\mathbf{w}_t)\|_2^2.$$

- If $\eta_t \in]0, \frac{2}{\mathsf{L}}[$ and $\nabla f(\mathbf{w}_t) \neq \mathbf{0}$, *strictly* decrease function value
- Rearranging:

$$\|\nabla f(\mathbf{w}_t)\|_2^2 \le \frac{f(\mathbf{w}_t) - f(\mathbf{w}_{t+1})}{\eta_t(1 - \mathsf{L}\eta_t/2)} \le \frac{f(\mathbf{w}_t) - f(\mathbf{w}_{t+1})}{\alpha\beta\mathsf{L}/2}.$$

- Telescoping:

$$\sum_{t=0}^{T-1} \|\nabla f(\mathbf{w}_t)\|_2^2 \le \frac{f(\mathbf{w}_0) - f(\mathbf{w}_T)}{\alpha\beta\mathsf{L}/2} \le \frac{f(\mathbf{w}_0) - f_\star}{\alpha\beta\mathsf{L}/2}.$$

Remarkable Properties

- Rate of convergence is proportional to the Lipschitz smoothness L: the bigger L is, the smaller the step size $\eta = \frac{1}{\mathsf{L}}$ has to be since the function f becomes steeper.
- If we start from some point $\mathbf{w}_0$ whose function value is closer to the infimum $f_\star$, then the gradient diminishes faster to zero.
- Very importantly, the rate of convergence does not depend on d, the dimension, at all!
- The $1/\sqrt{T}$ rate of convergence for the gradient is essentially tight[1].

[1] C. Cartis, N. I. M. Gould, and P. L. Toint. "On the Complexity of Steepest Descent, Newton's and Regularized Newton's Methods for Nonconvex Unconstrained Optimization". *SIAM Journal on Optimization*, vol. 20, no. 6 (2010), pp. 2833–2852.

Backtracking

- Figuring out L can be tedious; and it can be conservative too
- Where did we use the knowledge of L in the proof?

$$f(\mathbf{w}_{t+1}) = f(\mathbf{w}_t - \eta_t \nabla f(\mathbf{w}_t)) \leq f(\mathbf{w}_t) - \eta_t \underbrace{\left(1 - \frac{\mathsf{L}\eta_t}{2}\right)}_{\geq 0} \|\nabla f(\mathbf{w}_t)\|_2^2.$$

- Choose some $\alpha \in]0, 1[$, say $\alpha = \frac{1}{2}$, and aim:

$$f(\mathbf{w}_t - \eta_t \nabla f(\mathbf{w}_t)) \leq f(\mathbf{w}_t) - \alpha\eta_t \|\nabla f(\mathbf{w}_t)\|_2^2.$$

- The above inequality is testable without knowing L!
  - if the test succeeds, happily proceed to the next iteration
  - if the test fails, halve η_t and repeat
  - $\eta_t \geq \frac{1-\alpha}{\mathsf{L}}$, repeat at most $K := \log_2 \frac{\eta \mathsf{L}}{1-\alpha}$ times

L. Armijo. "Minimization of functions having Lipschitz continuous first partial derivatives". *Pacific Journal of Mathematics*, vol. 16, no. 1 (1966), pp. 1–3.