M2177.0043 Introduction to Deep Learning

Lecture 3: Unconstrained optimization[1]

Hyun Oh Song[1]

[1] Dept. of Computer Science and Engineering, Seoul National University

March 24, 2020

[1] Many slides and figures adapted from Stephen Boyd

Last time

- Linear algebra review

Outline

Mathematical optimization

$$\begin{aligned} \underset{x}{\text{minimize}} \quad & f(x) \\ \text{subject to} \quad & h_i(x) \leqslant b_i, \ \forall i = 1, \ldots, m \end{aligned} \tag{1}$$

- $x = (x_1, \ldots, x_n)$ is the *optimization variable* of the problem 1
- $f : \mathbb{R}^n \to \mathbb{R}$ is the *objective function*
- $h_i : \mathbb{R}^n \to \mathbb{R}$ is the *constraint function*
- x^* is called the *minimizer* or *solution* of the problem 1, if it has the smallest objective value among all vectors that satisfy the constraints.

Global and local minimum

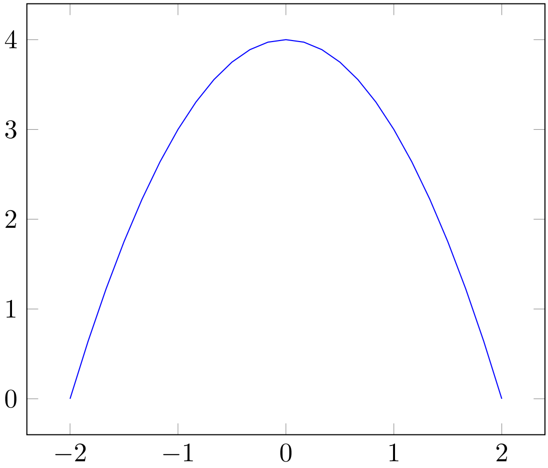
Definition 1 (Global minimum)

A real-valued function f defined on a domain $\mathcal{X}$ has a global minimum at x^* if $f(x^*) \leqslant f(x)\ \forall x \in \mathcal{X}$.

Definition 2 (Local minimum)

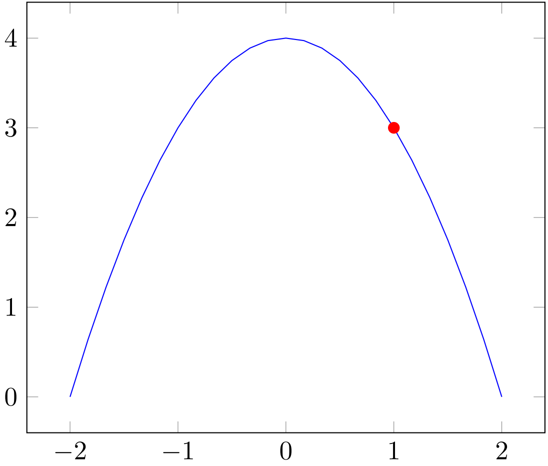
A real-valued function f defined on a domain $\mathcal{X}$ has a local minimum at x^* if $\exists \epsilon > 0$ such that $f(x^*) \leqslant f(x)\ \forall x \in \mathcal{X}$ within distance ϵ of x^*.

Example



$$
\begin{aligned}
\underset{x}{\text{minimize}} \quad & 4 - x^2 \\
\text{subject to} \quad & -x \leqslant 0 \\
& x \leqslant 1 \\
& x^2 \leqslant 2
\end{aligned}
$$

Example



$$\begin{aligned} \underset{x}{\text{minimize}} \quad & 4 - x^2 \\ \text{subject to} \quad & -x \leqslant 0 \\ & x \leqslant 1 \\ & x^2 \leqslant 2 \end{aligned}$$

minimizer $x^* = 1$,
minimum value $f(x^*) = 3$

Outline

Unconstrained minimization

$$\underset{x}{\text{minimize}}\, f(x)$$

- f differentiable
- assume optimal value $p^* = \inf_x f(x)$ is attained
- produce sequence of points $x^{(k)} \in \mathrm{dom} f,\ k = 0, 1, \ldots$

$$f(x^{(k)}) \to p^*$$

Descent methods

$$x^{(k+1)} = x^{(k)} + t^{(k)} \Delta x^{(k)} \quad \text{with } f(x^{(k+1)}) < f(x^{(k)})$$

- Δx is the *step*, or *search direction*
- t is the *step size*, or *step length*

General descent method.

given a starting point $x \in \mathrm{dom} f$.

repeat

1. Determine a descent direction Δx
2. *Line search.* Choose a step size $t > 0$.
3. *Update.* $x := x + t\Delta x$.

until stopping criterion is satisfied.

Line search

- exact line search
$$t = \operatorname*{argmin}_{t>0} f(x + t\Delta x)$$
- backtracking line search

Bisection line section method.

given a, b, ϵ

Set $A = a, B = b$

repeat

    if $f'(\frac{A+B}{2}) > 0$ **then** $B = \frac{A+B}{2}$
    else $A = \frac{A+B}{2}$
    end if

until $|B - A| \leqslant \epsilon$

Gradient descent method

general descent method with $\Delta x = -\nabla f(x)$

---

given a starting point $x \in \mathrm{dom} f$.
repeat

1. $\Delta x := -\nabla f(x)$.
2. *Line search.* Choose a step size $t > 0$.
3. *Update.* $x := x + t\Delta x$.

until stopping criterion is satisfied.

- stopping criterion usually of the form $||\nabla f(x)||_2 \leqslant \epsilon$
- very simple, but often very slow; rarely used in practice
- for convergence analysis, take my graduate level machine learning class.

Why is the negative gradient the direction of steepest descent?

- Consider the rate of change of function f at point $\mathbf{x} \in \operatorname{dom} f$ along a unit vector $\mathbf{v}$ pointing in an arbitrary direction. This is called the *directional derivative* of a function $D_{\mathbf{v}} f(\mathbf{x})$.

$$D_{\mathbf{v}} f(\mathbf{x}) = \lim_{h \to 0} \frac{f(\mathbf{x} + h\mathbf{v}) - f(\mathbf{x})}{h}$$

- So the question "What is the direction of steepest descent of f at $\mathbf{x}$?" can be translated to "For which $\mathbf{v}$, is $D_{\mathbf{v}} f(\mathbf{x})$ minimized?

- Now, it can be proven that if f is differentiable at $\mathbf{x}$, the limit above evaluates to $D_{\mathbf{v}} f(\mathbf{x}) = \nabla_{\mathbf{x}} f(\mathbf{x})^{\intercal} \mathbf{v}$ (proof in slide 15)

- Using the law of cosines / dot product, we know:

$$D_{\mathbf{v}} f(\mathbf{x}) = \|\nabla_{\mathbf{x}} f(\mathbf{x})\| \|\mathbf{v}\| \cos\theta = \|\nabla_{\mathbf{x}} f(\mathbf{x})\| \cos\theta$$

  where θ is the angle between the vectors $\nabla_{\mathbf{x}} f(\mathbf{x})$ and $\mathbf{v}$. Now $\cos\theta$ achieves minimum value of -1 when $\theta = \pi$. The angle between $\nabla_{\mathbf{x}} f(\mathbf{x})$ and $\mathbf{v}$ are actually pointing the opposite direction.

- That is: the minimum value of $D_{\mathbf{v}} f(\mathbf{x})$ is $\nabla_{\mathbf{x}} f(\mathbf{x})$ and is achieved when $\mathbf{v}$ points in $-\nabla_{\mathbf{x}} f(\mathbf{x})$.

Prove $D_{\mathbf{v}} f(\mathbf{x}) = \nabla_{\mathbf{x}} f(\mathbf{x})^\intercal \mathbf{v}$ if f is differentiable

We'll prove in two variable case but it's easy to generalize for n variables. Let $\mathbf{v} = (a, b)$, $g(t) = f(x_0 + ta, y_0 + tb) = f(x, y)$.

$$
\begin{aligned}
g'(t) &= \frac{\partial f}{\partial x}\frac{dx}{dt} + \frac{\partial f}{\partial y}\frac{dy}{dt} = \frac{\partial f}{\partial x} a + \frac{\partial f}{\partial y} b \\
g'(0) &= \frac{\partial f(x_0, y_0)}{\partial x} a + \frac{\partial f(x_0, y_0)}{\partial y} b = \nabla f(x_0, y_0)^\intercal \mathbf{v}
\end{aligned}
$$

Also, $g'(0)$ is identical to the directional derivative of f at point (x_0, y_0) along direction $\mathbf{v}$ because,

$$
g'(t) = \lim_{h \to 0} \frac{g(t+h) - g(t)}{h}, \qquad g'(0) = \lim_{h \to 0} \frac{g(h) - g(0)}{h}
$$

$$
g'(0) = \lim_{h \to 0} \frac{f(x_0 + ha, y_0 + hb) - f(x_0, y_0)}{h} = D_{\mathbf{v}} f(x_0, y_0)
$$

$\therefore D_{\mathbf{v}} f(x_0, y_0) = \nabla f(x_0, y_0)^\intercal \mathbf{v}$

Quadratic problem in $\mathbb{R}^2$

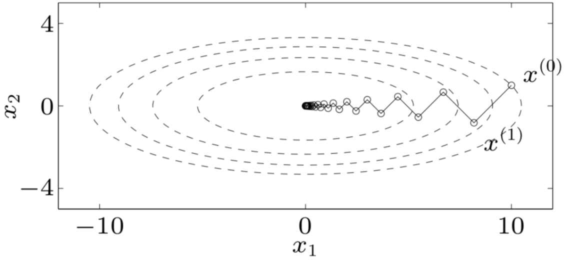
$$f(x) = \frac{1}{2}(x_1^2 + \gamma x_2^2) \qquad (\gamma > 0)$$

with exact line search, starting at $x^{(0)} = (\gamma, 1)$:

$$x_1^{(k)} = \gamma \left(\frac{\gamma - 1}{\gamma + 1}\right)^k, \qquad x_2^{(k)} = \left(-\frac{\gamma - 1}{\gamma + 1}\right)^k \qquad (\textbf{why?})$$

- what is the solution? very slow if $\gamma \gg 1$ or $\gamma \ll 1$

- example for $\gamma = 10$:



- How many steps does it take to converge if $\gamma = 1$?

Newton step I

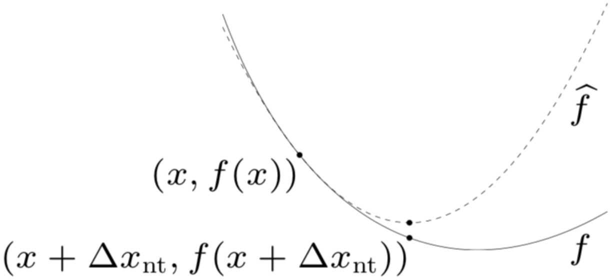
$$\Delta x_{\mathsf{nt}} = -\nabla^2 f(x)^{-1} \nabla f(x)$$

interpretations

- $x + \Delta x_{\mathsf{nt}}$ minimizes second order approximation

$$\widehat{f}(x+v) = f(x) + \nabla f(x)^{\mathsf{T}} v + \frac{1}{2} v^{\mathsf{T}} \nabla^2 f(x) v$$

Newton step II

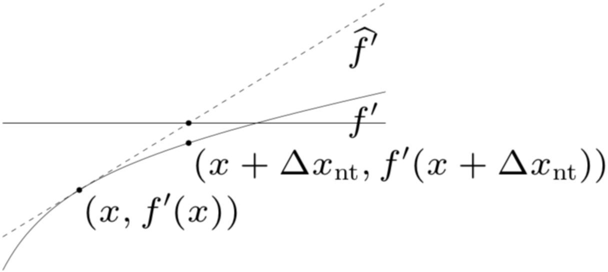


Newton step III

- $x + \Delta x_{\mathrm{nt}}$ solves linearized optimality condition. Find the direction where the gradient evaluated at $x + v$ is zero.

$$\nabla f(x+v) = 0$$

$$\nabla \widehat{f}(x+v) = \nabla f(x) + \nabla^2 f(x) v = 0 \qquad \text{linearized approximation } \widehat{f}$$



Newton's method

---

given a starting point $x \in \mathrm{dom} f$.

repeat

1. *Compute the Newton step* $\Delta x_{\mathsf{nt}} := -\nabla^2 f(x)^{-1} \nabla f(x)$.
2. *Line search.* Choose a step size $t > 0$.
3. *Update.* $x := x + t\Delta x_{\mathsf{nt}}$.

until stopping criterion is satisfied.

---

affine invariant, *i.e.*, independent of linear change of coordinates

Outline

Coordinate descent

- Gradient descent method updates **all** variables simultaneously with gradient

- Coordinate descent
  - update **one** variable at a time
  - may or may not use the gradient; sometimes analytically solvable in one variable

Coordinate descent

---

given a starting point $x^{(0)} \in \mathrm{dom} f, k := 0$.

repeat

1. $x_1^{(k)} := \mathrm{argmin}_{x_1} f(x_1, x_2^{(k-1)}, \ldots, x_n^{(k-1)})$.
2. $x_2^{(k)} := \mathrm{argmin}_{x_2} f(x_1^{(k)}, x_2, \ldots, x_n^{(k-1)})$.

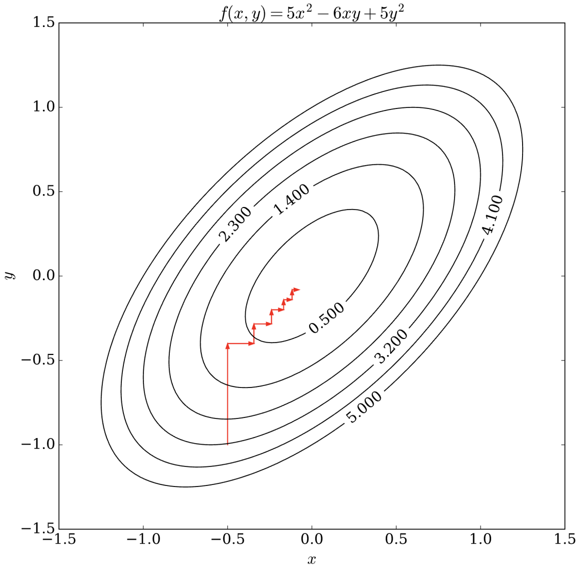
$\vdots$

n. $x_n^{(k)} := \mathrm{argmin}_{x_n} f(x_1^{(k)}, x_2^{(k)}, \ldots, x_n)$.

$k := k + 1$

until stopping criterion is satisfied.

Coordinate descent



Example: coordinate descent on convex function

Show that coordinate descent fails for the function g. Verify that the algorithm terminates after one step at the point $(x_2^{(0)}, x_2^{(0)})$, while $\inf_x g(x) = -\infty$. (To see this, set $x = (-t, -t)$ and let $t \to \infty$, we see that $g(x) = -0.2t \to -\infty$)

$$g(x) = |x_1 - x_2| + 0.1(x_1 + x_2)$$

Example: coordinate descent on convex function

Show that coordinate descent fails for the function g. Verify that the algorithm terminates after one step at the point $(x_2^{(0)}, x_2^{(0)})$, while $\inf_x g(x) = -\infty$. (To see this, set $x = (-t, -t)$ and let $t \to \infty$, we see that $g(x) = -0.2t \to -\infty$)

$$g(x) = |x_1 - x_2| + 0.1(x_1 + x_2)$$

solution

First minimize over x_1 with x_2 fixed as $x_2^{(0)}$. w.l.o.g, assume $x_1 > x_2$, $f(x_1) = 1.1x_1 - 0.9x_2^{(0)}$. Optimal $x_1 = x_2^{(0)}$. We arrive at $(x_2^{(0)}, x_2^{(0)})$. We now optimize over x_2 but it is optimal by symmetry, so x is unchanged. We're now at a fixed point of the coordinate-descent algorithm. Even though f is convex, coordinate descent does not guarantee global minima.