Lecture 2: Gradient Methods

- Gradient Methods - Motivation
- Principal Gradient Methods
- Gradient Methods - Choices of Direction
- Asymptotic Convergence
- Local Convergence Rate

Motivation

Consider the minimization of a differentiable function f

$$\begin{array}{ll} \text{minimize} & f(\boldsymbol{x}) \\ \text{subject to} & \boldsymbol{x} \in \mathbb{R}^n \end{array}$$

Suppose we are at a point $\boldsymbol{x}$.

- If $\nabla f(\boldsymbol{x}) = \mathbf{0}$, then $\boldsymbol{x}$ is a candidate solution. Done.
- If $\nabla f(\boldsymbol{x}) \neq \mathbf{0}$, there is an interval $(0, \delta)$ of stepsizes such that

$$f(\boldsymbol{x} - \alpha \nabla f(\boldsymbol{x})) < f(\boldsymbol{x}), \quad \text{for all } \alpha \in (0, \delta).$$

- Moreover, if the direction $\boldsymbol{d}$ makes an angle with $\nabla f(\boldsymbol{x})$ that is greater than 90 degrees,

$$\nabla f(\boldsymbol{x})' \boldsymbol{d} < 0,$$

there is an interval $(0, \delta)$ of stepsizes such that

$$f(\boldsymbol{x} + \alpha \boldsymbol{d}) < f(\boldsymbol{x}) \quad \text{for all } \alpha \in (0, \delta).$$

Iterative Descent Methods

$$\boxed{\boldsymbol{x}^{r+1} = \boldsymbol{x}^r + \alpha_r d^r, \quad r = 0, 1, ...}$$

where, if $\nabla f(\boldsymbol{x}^r) \neq \mathbf{0}$, the direction $\boldsymbol{d}^r$ satisfies $\nabla f(\boldsymbol{x}^r)'\boldsymbol{d}^r < 0$, and α_r is a positive stepsize.

- Principal example:

$$\boldsymbol{x}^{r+1} = \boldsymbol{x}^r - \alpha_r \boldsymbol{D}^r \nabla f(\boldsymbol{x}^r), \quad r = 0, 1, ...$$

  where $\boldsymbol{D}^r$ is a positive definite symmetric matrix

- Simplest method: Steepest descent

$$\boldsymbol{x}^{r+1} = \boldsymbol{x}^r - \alpha_r \nabla f(\boldsymbol{x}^r), \quad r = 0, 1, ...$$

- Most sophisticated method: Newton's method

$$\boldsymbol{x}^{r+1} = \boldsymbol{x}^r - \alpha_r \nabla^2 f(\boldsymbol{x}^r)^{-1} \nabla f(\boldsymbol{x}^r), \quad r = 0, 1, ...$$

Discussion

- Steepest descent: no memory, slow convergence
  - ⋆ zigzagging
  - ⋆ consider the convex quadratic case; convergence related to the condition number. Plot.
- Newton's method: no memory, fast convergence (with $\alpha_r = 1$);
  - ⋆ Given $\boldsymbol{x}^r$, the method obtains $\boldsymbol{x}^{r+1}$ as the minimum of a quadratic approximation of f based on a second order Taylor expansion around $\boldsymbol{x}^r$. Plot.
  - ⋆ insensitive to condition number
  - ⋆ how many Newton iterations does it take to minimize a quadratic function f?

Other Choices of Direction

- Diagonally Scaled Steepest Descent:

$$\boldsymbol{D}^r = \text{Diagonal approximation to } \left(\nabla^2 f(\boldsymbol{x}^r)\right)^{-1}.$$

- Modified Newton's Method:

$$\boldsymbol{D}^r = \left(\nabla^2 f(\boldsymbol{x}^0)\right)^{-1}, \quad r = 0, 1, ...$$

- Discretized Newton's Method:

$$\boldsymbol{D}^r = \left(\boldsymbol{H}(\boldsymbol{x}^r)\right)^{-1}, \quad r = 0, 1, ...,$$

where $\boldsymbol{H}(\boldsymbol{x}^r)$ is a finite-difference based approximation of $\nabla^2 f(\boldsymbol{x}^r)$.

- Gauss-Newton method for least squares problems $\min_{\boldsymbol{x}} \frac{1}{2}\|\boldsymbol{g}(\boldsymbol{x})\|^2$. Here $\boldsymbol{D}^r$ is an approximation of the true Hessian inverse $\left(\nabla^2 \boldsymbol{g}(\boldsymbol{x}^r)\right)^{-1}$

$$\boldsymbol{D}^r = \left(\nabla \boldsymbol{g}(\boldsymbol{x}^r) \nabla \boldsymbol{g}(\boldsymbol{x}^r)'\right)^{-1}, \quad r = 0, 1, ...$$

where $\nabla \boldsymbol{g}(\boldsymbol{x}^r)$ is the Jacobian matrix of $\boldsymbol{g}$.

Choices of Stepsize

- **Minimization Rule:** α_r is such that

$$f(\boldsymbol{x}^r + \alpha_r \boldsymbol{d}^r) = \min_{\alpha \geq 0} f(\boldsymbol{x}^r + \alpha \boldsymbol{d}^r).$$

- **Limited Minimization Rule:** Fix some $s > 0$. Choose α_r such that

$$f(\boldsymbol{x}^r + \alpha_r \boldsymbol{d}^r) = \min_{\alpha \in [0,s]} f(\boldsymbol{x}^r + \alpha \boldsymbol{d}^r).$$

- **Constant stepsize:** α_r is such that $\alpha_r = s$ is a constant

- **Diminishing stepsize:** $\alpha_r \to 0$ but satisfies the infinite travel condition

$$\sum_{r=0}^{\infty} \alpha_r = \infty$$

- **Armijo rule:** Let $\sigma \in (0, \frac{1}{2})$. Start with s and continue with $\beta s,\ \beta^2 s, \ldots,$ until $\beta^m s$ falls within the set of α with

$$f(\boldsymbol{x}^r) - f(\boldsymbol{x}^r + \alpha \boldsymbol{d}^r) \geq -\sigma \alpha \nabla f(\boldsymbol{x}^r)' \boldsymbol{d}^r.$$

Claim: if $\boldsymbol{d}^r = -\boldsymbol{D}^r \nabla f(\boldsymbol{x}^r) \neq \boldsymbol{0}$ with $\boldsymbol{D}^r \succ \boldsymbol{0}$, then m is finite.

Convergence of Iterative Methods

- **Convergence to stationary points**
  - ⋆ Sanity check
  - ⋆ Minimal requirement of any reasonable algorithm
  - ⋆ Does not give global efficiency of the algorithm
- **Asymptotic convergence rate:** local analysis, assuming already close to a solution, let # of iterations go to infinity
  - ⋆ **1) linear, 2) supperlinear, 3) sublinear**
- **Iteration complexity analysis:**
  - ⋆ Measures the number of iterations required to get an ϵ optimal solution (e.g., $f(\boldsymbol{x}^r) - f^* \leq \epsilon$)
  - ⋆ Current analysis is all for the worst case and requires convexity
  - ⋆ Gives global behavior of the algorithm

Convergence of Gradient Descent Methods

- Only convergence to stationary points can be guaranteed
- Even convergence to a single limit may be hard to guarantee (capture theorem)
- Danger of nonconvergence if directions $\boldsymbol{d}^r$ tend to be orthogonal to $\nabla f(\boldsymbol{x}^r)$
- **Gradient related condition:** For any subsequence $\{\boldsymbol{x}^r\}_{k\in\mathcal{K}}$ that converges to a nonstationary point, the corresponding subsequence $\{\boldsymbol{d}^r\}_{r\in\mathcal{K}}$ is bounded and satisfies

$$\limsup_{r\to\infty, r\in\mathcal{K}} \nabla f(\boldsymbol{x}^r)'\boldsymbol{d}^r < 0.$$

- The condition is satisfied if $\boldsymbol{d}^r = -\boldsymbol{D}^r\nabla f(\boldsymbol{x}^r)$ and the eigenvalues of $\boldsymbol{D}^r$ are bounded above and bounded away from zero

Convergence Issues

- Let $\{\boldsymbol{x}^r\}$ be a sequence generated by a gradient method

$$\boldsymbol{x}^{r+1} = \boldsymbol{x}^r + \alpha_r \boldsymbol{d}^r,$$

where $\{\boldsymbol{d}^r\}$ is gradient related. Assume that for some constant $L > 0$, we have

$$\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\| \leq L\|\boldsymbol{x} - \boldsymbol{y}\|, \forall\ \boldsymbol{x},\ \boldsymbol{y} \in \mathbb{R}^n.$$

Assume that either (a) there exists a scalar ϵ such that for all r

$$0 < \epsilon \leq \alpha_r \leq \frac{(2-\epsilon)|\nabla f(\boldsymbol{x}^r)'\boldsymbol{d}^r|}{L\|\boldsymbol{d}^r\|^2} \tag{1}$$

or (b) $\alpha_r \to 0$ and

$$\sum_{r=1}^{\infty} \alpha_r = \infty.$$

Then either $f(\boldsymbol{x}^r) \to -\infty$ or else $\{f(\boldsymbol{x}^r)\}$ converges to a finite value and $\nabla f(\boldsymbol{x}^r) \to \boldsymbol{0}$.

- Specialize to the steepest descent: $\boldsymbol{d}^r = -\nabla f(\boldsymbol{x}^r)$.

Convergence Analysis

- Given $\boldsymbol{x}^r$ and the descent direction $\boldsymbol{d}^r$, the Lipschitz assumption implies

$$f(\boldsymbol{x}^r + \alpha \boldsymbol{d}^r) - f(\boldsymbol{x}^r) \leq \alpha \nabla f(\boldsymbol{x}^r)' \boldsymbol{d}^r + \frac{1}{2}\alpha^2 L \|\boldsymbol{d}^r\|^2$$

- Minimization of this function over α yields the stepsize

$$\bar{\alpha}_r = -\frac{\nabla f(\boldsymbol{x}^r)' \boldsymbol{d}^r}{L\|\boldsymbol{d}^r\|^2}$$

- In case of steepest descent, $\bar{\alpha}_r = 1/L$.

- The constant stepsize (1) reduces the cost function f at each iteration.

$$f(\boldsymbol{x}^r + \alpha_r \boldsymbol{d}^r) - f(\boldsymbol{x}^r) \leq \alpha_r \mu \nabla f(\boldsymbol{x}^r)' \boldsymbol{d}^r \tag{2}$$

  for some constant $\mu > 0$.

Convergence Analysis, Continued

- Assume $\bar{\boldsymbol{x}}$ is a nonstationary limit point. Then $f(\boldsymbol{x}^r) \downarrow f(\bar{\boldsymbol{x}})$, it follows from (2) that

$$\alpha_r \nabla f(\boldsymbol{x}^r)' \boldsymbol{d}^r \to 0.$$

- If $\{\boldsymbol{x}^r\}_{r\in\mathcal{K}} \to \bar{\boldsymbol{x}}$, then by the gradient relatedness

$$\limsup_{r\to\infty, r\in\mathcal{K}} \nabla f(\boldsymbol{x}^r)' \boldsymbol{d}^r < 0.$$

- This implies $\{\alpha_r\}_{r\in\mathcal{K}} \to 0$, contradicting stepsize (1).

- For the diminishing stepsize rule (b), assume the whole sequence converges to a non-stationary point $\bar{\boldsymbol{x}}$. Then the gradient relatedness implies

$$\limsup_{r\to\infty} \nabla f(\boldsymbol{x}^r)' \boldsymbol{d}^r < 0.$$

Sum (2) over all r shows $f(\boldsymbol{x}^r) \to -\infty$, contradiction.

- Similar proof for the Armijo rule. For large $r \in \mathcal{K}$

$$f(\boldsymbol{x}^r) - f\left(\boldsymbol{x}^r + (\alpha_r/\beta)\boldsymbol{d}^r\right) < -\sigma\frac{\alpha_r}{\beta}\nabla f(\boldsymbol{x}^r)'\boldsymbol{d}^r.$$

- Defining $\boldsymbol{p}^r = \boldsymbol{d}^r/\|\boldsymbol{d}^r\|$ and $\bar{\alpha}_r = \alpha_r\|\boldsymbol{d}^r\|/\beta$, we have

$$\frac{f(\boldsymbol{x}^r) - f(\boldsymbol{x}^r + \bar{\alpha}_r\boldsymbol{p}^r)}{\bar{\alpha}_r} < -\sigma\nabla f(\boldsymbol{x}^r)'\boldsymbol{p}^r.$$

- Letting $k \to \infty$, we get

$$-\nabla f(\bar{\boldsymbol{x}})'\bar{\boldsymbol{p}} \leq -\sigma\nabla f(\bar{\boldsymbol{x}})'\bar{\boldsymbol{p}},$$

  where $\bar{\boldsymbol{p}}$ is a limit point of $\{\boldsymbol{p}^r\}_{r\in\mathcal{K}}$.

- This is a contradiction since $\nabla f(\bar{\boldsymbol{x}})'\bar{\boldsymbol{p}} < 0$.

Other Stepsize Rules

- **Non-monotone step size rule:** Fix some integer $K > 0$. Reduce α by a factor of β until

$$f(\boldsymbol{x}^r + \alpha \boldsymbol{d}^r) - \max_{k=1,2,\ldots,K} f(\boldsymbol{x}^{r-k}) \leq \sigma \alpha \nabla f(\boldsymbol{x}^r)' \boldsymbol{d}^r.$$

  The resulting sequence $\{f(\boldsymbol{x}^r)\}$ is no longer monotonically decreasing. But the convergence to stationary point can still be established.

- **BB (Barzilai-Borwein) stepsize rule:** consider minimizing a strongly convex quadratic function

$$\text{minimize}_{\boldsymbol{x} \in \mathbb{R}^n} \quad f(\boldsymbol{x}) = \frac{1}{2} \boldsymbol{x}' \boldsymbol{A} \boldsymbol{x} + \boldsymbol{b}' \boldsymbol{x}$$

The steepest descent method with exact line minimization stepsize rule is

$$\alpha_r = \frac{\|\nabla f(\boldsymbol{x}^r)\|^2}{(\nabla f(\boldsymbol{x}^r))' \boldsymbol{A} \nabla f(\boldsymbol{x}^r)}.$$

The BB stepsize rule uses

$$\alpha_r^{\mathrm{BB}} = \frac{\|\nabla f(\boldsymbol{x}^{r-1})\|^2}{(\nabla f(\boldsymbol{x}^{r-1}))' \boldsymbol{A} \nabla f(\boldsymbol{x}^{r-1})}.$$

Practical performance of BB stepsize rule is quite good (compared to the steepest descent) in some applications, but the theoretical understanding of BB method remains limited (more later).

Local Convergence Rate Analysis

- Restrict attention to sequences $\{\boldsymbol{x}^r\}$ converging to a local min $\boldsymbol{x}^*$.
- Measure progress in terms of an error function $e(\boldsymbol{x})$ with $e(\boldsymbol{x}^*) = 0$, such as

$$e(\boldsymbol{x}) = \|\boldsymbol{x} - \boldsymbol{x}^*\|, \quad e(\boldsymbol{x}) = f(\boldsymbol{x}) - f(\boldsymbol{x}^*)$$

- Compare the tail of the sequence $e(\boldsymbol{x}^r)$ with the tail of standard sequences.
- Geometric or linear convergence [if $e(\boldsymbol{x}^r) \leq q\beta^r$ for some $q > 0$ and $\beta \in [0, 1)$, and for all r]. Holds if

$$\limsup_{r\to\infty} \frac{e(\boldsymbol{x}^{r+1})}{e(\boldsymbol{x}^r)} < \beta$$

- Superlinear convergence [if $e(\boldsymbol{x}^r) \leq q\beta^{p^r}$ for some $q > 0,\ p > 1$ and $\beta \in [0, 1)$, and for all r].

Quadratic Model Analysis

- Focus on the quadratic function $f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}'\boldsymbol{Q}\boldsymbol{x}$, with $\boldsymbol{Q} \succ \boldsymbol{0}$. Optimal solution $\boldsymbol{x}^* = ?$
- Analysis also applies to nonquadratic problems in the neighborhood of a nonsingular local min. Consider steepest descent

$$\boldsymbol{x}^{r+1} = \boldsymbol{x}^r - \alpha_r \nabla f(\boldsymbol{x}^r) = (\boldsymbol{I} - \alpha_r \boldsymbol{Q})\boldsymbol{x}^r$$

$$\Rightarrow \|\boldsymbol{x}^{r+1}\|^2 = (\boldsymbol{x}^r)'(\boldsymbol{I} - \alpha_r \boldsymbol{Q})^2 \boldsymbol{x}^r \leq \max \text{eig.}(\boldsymbol{I} - \alpha_r \boldsymbol{Q})^2 \|\boldsymbol{x}^r\|^2$$

- The eigenvalues of $(\boldsymbol{I} - \alpha_r \boldsymbol{Q})^2$ are equal to $(1 - \alpha_r \lambda_i)^2$, where λ_i are the eigenvalues of $\boldsymbol{Q}$, so

$$\max \text{eig of } (\boldsymbol{I} - \alpha_r \boldsymbol{Q})^2 = \max\{(1 - \alpha_r m)^2, (1 - \alpha_r M)^2\}$$

where m, M are the smallest and largest eigenvalues of Q. Thus

$$\frac{\|\boldsymbol{x}^{r+1}\|}{\|\boldsymbol{x}^r\|} \leq \max\{|1 - \alpha_r m|, |1 - \alpha_r M|\}$$

Optimal Convergence Rate

- . The value of α_r that minimizes the bound is $\alpha^* = 2/(M+m)$, in which case

$$\frac{\|\boldsymbol{x}^{r+1}\|}{\|\boldsymbol{x}^r\|} \leq \frac{M-m}{M+m}$$

- Convergence rate of f for the minimization stepsize

$$\frac{f(\boldsymbol{x}^{r+1})}{f(\boldsymbol{x}^r)} \leq \left(\frac{M-m}{M+m}\right)^2$$

- The ratio $\kappa := M/m$ is called the condition number of Q, and problems with M/m: large are called ill-conditioned.
- Convergence rate is independent of problem dimension.
- **Iteration complexity:** find an ϵ-optimal solution, we need

$$O(\kappa \ln D/\epsilon)$$

  iterations, where $D := \|\boldsymbol{x}^0 - \boldsymbol{x}^*\|^2$.
- Converges in **one** step if $M = m$.

Causes of Ill-Conditioning

- Ill-conditioning usually arises when optimization variables are scaled incoherently.
- Example:
  - $\star$ x_1: gas flow, lbs/hr, nominal value $11,000$
  - $\star$ x_2: waste buildup: lbs, nominal value 6×10^{-4}.
  - $\star$ x_3: steam thermal resistance, BTU/(hr·ft^2·^{o}F)$^{-1}$; nominal value 100
- Should rescale the units so that the nominal values are balanced.

Scaled Steepest Descent

- View the more general method

$$\boldsymbol{x}^{r+1} = \boldsymbol{x}^r - \alpha_r \boldsymbol{D}^r \nabla f(\boldsymbol{x}^r)$$

  as a scaled version of steepest descent.
- Consider a change of variables $\boldsymbol{x} = \boldsymbol{S}\boldsymbol{y}$ with $\boldsymbol{S} = (\boldsymbol{D}^r)^{1/2}$. In the space of $\boldsymbol{y}$, the problem is

$$\begin{array}{ll} \text{minimize} & h(\boldsymbol{y}) := f(\boldsymbol{S}\boldsymbol{y}) \\ \text{subject to} & \boldsymbol{y} \in \mathbb{R}^n \end{array}$$

- Apply steepest descent to this problem, multiply with $\boldsymbol{S}$, and pass back to the space of $\boldsymbol{x}$, using $\nabla h(\boldsymbol{y}^r) = \boldsymbol{S}\nabla f(\boldsymbol{x}^r)$,

$$\begin{aligned} \boldsymbol{y}^{r+1} &= \boldsymbol{y}^r - \alpha_r \nabla h(\boldsymbol{y}^r) \\ \boldsymbol{S}\boldsymbol{y}^{r+1} &= \boldsymbol{S}\boldsymbol{y}^r - \alpha_r \boldsymbol{S}\nabla h(\boldsymbol{y}^r) \\ \boldsymbol{x}^{r+1} &= \boldsymbol{x}^r - \alpha_r \boldsymbol{D}^r \nabla f(\boldsymbol{x}^r) \end{aligned}$$

Diagonal Scaling

- Apply the results for steepest descent to the scaled iteration $\boldsymbol{y}^{r+1} = \boldsymbol{y}^r - \alpha_r \nabla h(\boldsymbol{y}^r)$:

$$\frac{\|\boldsymbol{y}^{r+1}\|}{\|\boldsymbol{y}^r\|} \leq \max\{|1-\alpha_r m_r|, |1-\alpha_r M_r|\}$$

$$\frac{f(\boldsymbol{x}^{r+1})}{f(\boldsymbol{x}^r)} = \frac{h(\boldsymbol{y}^{r+1})}{h(\boldsymbol{y}^r)} \leq \left(\frac{M_r - m_r}{M_r + m_r}\right)^2$$

where m_r and M_r are the smallest and largest eigenvalues of the Hessian of h, which is

$$\nabla^2 h(\boldsymbol{y}) = \boldsymbol{S}\nabla^2 f(\boldsymbol{x})\boldsymbol{S} = (\boldsymbol{D}^r)^{1/2}\boldsymbol{Q}(\boldsymbol{D}^r)^{1/2}$$

- It is desirable to choose $\boldsymbol{D}^r$ as close as possible to $\boldsymbol{Q}^{-1}$. Also if $\boldsymbol{D}^r$ is so chosen, the stepsize $\alpha = 1$ is near the optimal $2/(M_r + m_r)$.
- Using as $\boldsymbol{D}^r$ a diagonal approximation to $\boldsymbol{Q}^{-1}$ is common and often very effective. Corrects for poor choice of units expressing the variables.

Non-quadratic Problems

- Rate of convergence to a nonsingular local minimum of a non-quadratic function is very similar to the quadratic case (linear convergence is typical).

- If $\boldsymbol{D}^r = \left(\nabla^2 f(\boldsymbol{x}^*)\right)^{-1}$, we asymptotically obtain optimal scaling and superlinear convergence

- More generally, if the direction $\boldsymbol{d}^r = -\boldsymbol{D}^r \nabla f(\boldsymbol{x}^r)$ approaches asymptotically the Newton direction, i.e.,

$$\lim_{r\to\infty} \frac{\|\boldsymbol{d}^r + \left(\nabla^2 f(\boldsymbol{x}^*)\right)^{-1} \nabla f(\boldsymbol{x}^r)\|}{\|\nabla f(\boldsymbol{x}^r)\|} = 0$$

  and the Armijo rule is used with initial stepsize equal to one, the rate of convergence is superlinear.

- Convergence rate to a singular local min is typically sublinear (in effect, condition number $\kappa = \infty$)

Dealing with Nonnegative Constraints

$$\begin{array}{ll} \text{minimize} & f(\boldsymbol{x}) \\ \text{subject to} & \boldsymbol{x} \geq \mathbf{0}. \end{array}$$

- Convert to an unconstrained optimization problem: $x_i = w_i^2, \ i = 1, 2, ..., m$.
- Caution: this can create spurious solutions.
- Consider the example

$$\begin{array}{ll} \text{minimize} & (x_1 - 1)^2 + (x_2 + 1)^2 \\ \text{subject to} & x_1 \geq 0. \end{array}$$

- Optimal solution $\boldsymbol{x}^* = (1, -1)^T$. Conversion to an unconstrained problem:

$$\boxed{\begin{array}{ll} \text{minimize} & (w_1^2 - 1)^2 + (w_2 + 1)^2 \\ \text{subject to} & \boldsymbol{w} \in \mathbb{R}^2. \end{array}}$$

- We have

$$\nabla f = \begin{pmatrix} 4w_1(w_1^2 - 1) \\ 2(w_2 + 1) \end{pmatrix}, \quad \nabla^2 f = \begin{bmatrix} 4(3w_1^2 - 1) & 0 \\ 0 & 2 \end{bmatrix}$$

  which introduces a saddle point $(0, -1)^T$.

- We will discuss how to deal with inequality constraints later.