Distributed Optimization for Machine Learning

School of Electrical and Computer Engineering

University of Tehran

Erfan Darzi

Lecture 3 – Iterative Descent Methods and Convergence Analysis

erfandarzi@ut.ac.ir

Iterative Descent Methods

$$\min_{\mathbf{x}} \quad f(\mathbf{x})$$
$$\text{s.t.} \quad \mathbf{x} \in \mathbb{R}^n$$

- If $\nabla f(\mathbf{x}) = \mathbf{0}$, we have a candidate
- If $\nabla f(\mathbf{x}) \neq \mathbf{0}$, not a candidate ➔ Can we locally improve?

If $\nabla f(\mathbf{x})^T \mathbf{d} < \mathbf{0}$

$$\exists\, \delta > 0, \text{ s.t. } f(\mathbf{x} + \alpha \mathbf{d}) < f(\mathbf{x}), \quad \forall \alpha \in (0, \delta)$$

How to select

Choices of Direction

$$\mathbf{x}^{r+1} \leftarrow \mathbf{x}^r + \alpha^r \mathbf{d}^r$$

- Steepest/gradient descent: $\mathbf{d}^r = -\nabla f(\mathbf{x}^r)$
- Diagonally scaled gradient descent: $\mathbf{d}^r = -\mathbf{D}^r \nabla f(\mathbf{x}^r),$ for some $\mathbf{D}^r \succ \mathbf{0}$
- Newton direction (why?): $\mathbf{d}^r = -\left(\nabla^2 f(\mathbf{x}^r)\right)^{-1} \nabla f(\mathbf{x}^r)$
  - Benefit
  - Drawback

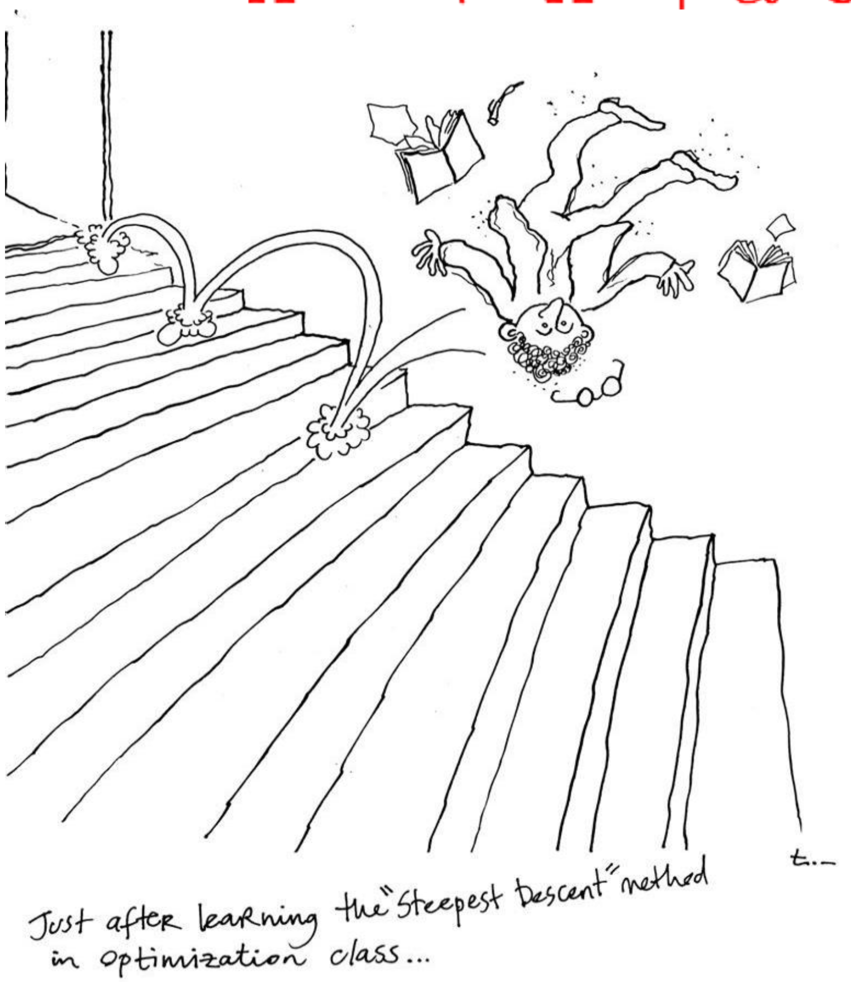
$$\mathbf{D}^r = \mathrm{diag}\left(\nabla^2 f(\mathbf{x}^r)\right)^{-1}$$

Choices of Step-size:

- Constant: $\alpha^r = \alpha, \ \forall r = 0, 1, \ldots$

$$\mathbf{x}^{r+1} \leftarrow \mathbf{x}^r + \alpha^r \mathbf{d}^r$$

Need to be careful about step-size!!



http://www.eurasip.org/DSPHumour/steepest-descent.jpg

Choices of Step-size:

$$\mathbf{x}^{r+1} \leftarrow \mathbf{x}^r + \alpha^r \mathbf{d}^r$$

- Constant: $\alpha^r = \alpha, \ \forall r = 0, 1, \ldots$
- Exact Minimization: $\alpha^r \in \arg\min_{\alpha \geq 0} f(\mathbf{x}^r + \alpha \mathbf{d}^r)$
- Limited Minimization $\alpha^r \in \arg \min_{\alpha \in (0, \bar{\alpha}]} f(\mathbf{x}^r + \alpha \mathbf{d}^r)$
- Diminishing: $\alpha^r \downarrow 0, \ \text{with} \sum_r \alpha^r = \infty$ → Why?
- Back-tracking/Armijo: Constants $\beta, \sigma \in (0,1)$ and initial stepsize $\bar{\alpha}$

$$\alpha^r = \max\{\bar{\alpha}\beta^i \mid f(\mathbf{x}^r) - f(\mathbf{x}^r + \bar{\alpha}\beta^i \mathbf{d}^r) \geq -\sigma \bar{\alpha}\beta^i \nabla f(\mathbf{x}^r)^T \mathbf{d}^r, \ i = 0, 1, \ldots\}$$

Actual decrease — Predicted decrease

Claim: If $\langle \nabla f(\mathbf{x}^r), \mathbf{d}^r \rangle < \mathbf{0}$, then α^r is well-defined

Convergence Analysis

Step-size + Direction $\rightarrow$ Algorithm

- Convergence to a stationary point (set of stationary points)
- Typical minimum requirement

- Asymptotic rate of convergence (**Convergence rate**) Assume $\{\mathbf{x}^r\} \rightarrow \mathbf{x}^*$
  - Error function examples: $e(\mathbf{x}) = \|\mathbf{x} - \mathbf{x}^*\|$ or $e(\mathbf{x}) = f(\mathbf{x}) - f(\mathbf{x}^*)$
  - Asymptotic behavior

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

$\beta \in (0,1):$ linear

$\beta = 1:$ sublinear

$\beta = 0:$ superlinear

- Iteration complexity analysis: Why we call it linear?
- Number of iterations required to achieve $\epsilon-$optimal solution: $e(\mathbf{x}^r) \leq \epsilon$
- Currently, worst case analysis

Convergence to Stationary Points

- To a single limit point may not be easy

- **Gradient related condition**: For any subsequence $\{\mathbf{x}^r\}_{r\in\mathcal{K}}$ converging to a non-stationary point, the corresponding subsequence is bounded and

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

- Example: $\mathbf{d}^r = -\mathbf{D}^r \nabla f(\mathbf{x}^r)$ with $\bar{\gamma}\mathbf{I} \succeq \mathbf{D}^r \succeq \underline{\gamma}\mathbf{I} \succ \mathbf{0}, \ \forall r$

Convergence to Stationary Points

- Assume:
  - $\mathbf{x}^{r+1} \leftarrow \mathbf{x}^r + \alpha^r \mathbf{d}^r$
  - $\mathbf{d}^r$ gradient related
  - Lipschitz gradient: $\exists L > 0 \text{ s.t. } \|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \leq L\|\mathbf{x} - \mathbf{y}\|, \ \ \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^n$
  - One of the following stepsize rules:
    - (a) Diminishing: $\alpha^r \to 0, \text{ and } \sum_r \alpha^r = \infty$
    - (b) Armijo
    - (c) Small enough: $0 < \epsilon \leq \alpha^r \leq \frac{(2-\epsilon)|\nabla f(\mathbf{x}^r)^T \mathbf{d}^r|}{L\|\mathbf{d}^r\|^2}$

- Then, every limit point of the iterates is a stationary point, i.e.,

$$\text{if } \{\mathbf{x}^r\}_{r \in \mathcal{K}} \to \bar{\mathbf{x}}, \ \text{then } \nabla f(\bar{\mathbf{x}}) = 0$$

- **Special case: gradient direction**
- **Proof (Requires descent lemma)** $f(\mathbf{x}+\mathbf{h}) \leq f(\mathbf{x}) + \mathbf{h}^T \nabla f(\mathbf{x}) + \frac{L}{2}\|\mathbf{h}\|^2$ → Why useful? → Proof
- **These are (asymptotically) monotone rules**

No assumption on convexity!