Gradient descent update is the minimizer of local quadratic over-estimator function at $\mathbf{x}_k$

Andersen Ang

Mathématique et recherche opérationnelle
UMONS, Belgium

manshun.ang@umons.ac.be Homepage: angms.science

First draft: June, 16, 2018
Last update : October 24, 2020

Gradient descent (GD)

- Unconstrained optimization problem of a differentiable function f

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

- As f is differentiable, gradient exists and GD update can be used

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha_k \nabla f(\mathbf{x}_k), \tag{1}$$

  where
  - $\alpha_k > 0$ is a (suitable) step size
  - $\nabla f(\mathbf{x}_k)$ is the gradient of f with respect to (w.r.t) $\mathbf{x}_k$
  - k is iteration counter, $k = 1, 2, ...$

- Question : where does the iteration (1) come from?

GD step as the minimizer to a local quadratic model

- The GD step actually is the minimizer to a function F:
  - F is a quadratic function of $\mathbf{x}$.
  - F is parameterized by $\mathbf{x}_k$.

- These two statements can be expressed mathematically as

$$\mathbf{x}_k - \alpha_k \nabla f(\mathbf{x}_k) = \underset{\mathbf{x}}{\operatorname{argmin}}\ F(\mathbf{x}; \mathbf{x}_k).$$

- Question is : what is such F?

  Answer : F is a local quadratic approximation of of f at $\mathbf{x}_k$ as

$$F(\mathbf{x}; \mathbf{x}_k) = f(\mathbf{x}_k) + \langle \nabla f(\mathbf{x}_k), \mathbf{x} - \mathbf{x}_k \rangle + \frac{1}{2\alpha_k} \|\mathbf{x} - \mathbf{x}_k\|_2^2.$$

A closer look at the local quadratic model ... (1/2)

- The local quadratic approximation of of f at $\mathbf{x}_k$:

$$F(\mathbf{x};\mathbf{x}_k) = f(\mathbf{x}_k) + \langle \nabla f(\mathbf{x}_k), \mathbf{x} - \mathbf{x}_k \rangle + \frac{1}{2\alpha_k}\|\mathbf{x} - \mathbf{x}_k\|_2^2.$$

- 1st-order Taylor approximation of f at $\mathbf{x}_k$

$$L(\mathbf{x};\mathbf{x}_k) = f(\mathbf{x}_k) + \langle \nabla f(\mathbf{x}_k), \mathbf{x} - \mathbf{x}_k \rangle.$$

- The local quadratic model $F(\mathbf{x};\mathbf{x}_k)$ can be viewed as the proximal regularization of a linearized model of f at the point $\mathbf{x}_k$

$$F(\mathbf{x};\mathbf{x}_k) := f(\mathbf{x}_k) + (\mathbf{x} - \mathbf{x}_k)^\top \nabla f(\mathbf{x}_k) + \frac{1}{2\alpha_k}\|\mathbf{x} - \mathbf{x}_k\|_2^2.$$

  - $f(\mathbf{x}_k) + (\mathbf{x} - \mathbf{x}_k)^\top \nabla f(\mathbf{x}_k)$ is a linear model of f at $\mathbf{x}_k$.
  - $\frac{1}{2\alpha_k}\|\mathbf{x} - \mathbf{x}_k\|_2^2$ is the proximal regularization term: it prevents $\mathbf{x}$ from moving too far away from $\mathbf{x}_k$.

A closer look at the local quadratic model ... (2/2)

- The local quadratic approximation of of f at $\mathbf{x}_k$:

$$\begin{aligned} F(\mathbf{x};\mathbf{x}_k) &= f(\mathbf{x}_k) + \langle \nabla f(\mathbf{x}_k), \mathbf{x}-\mathbf{x}_k\rangle + \frac{1}{2\alpha_k}\|\mathbf{x}-\mathbf{x}_k\|_2^2 \\ &= f(\mathbf{x}_k) + \langle \nabla f(\mathbf{x}_k), \mathbf{x}-\mathbf{x}_k\rangle + \frac{1}{2}\langle \frac{1}{\alpha_k}\mathbf{I}(\mathbf{x}-\mathbf{x}_k), \mathbf{x}-\mathbf{x}_k\rangle. \end{aligned}$$

- 2nd-order Taylor approximation of f at $\mathbf{x}_k$

$$H(\mathbf{x};\mathbf{x}_k) = f(\mathbf{x}_k) + \langle \nabla f(\mathbf{x}_k), \mathbf{x}-\mathbf{x}_k\rangle + \frac{1}{2}\langle \mathbf{H}(\mathbf{x}-\mathbf{x}_k), \mathbf{x}-\mathbf{x}_k\rangle,$$

  where $\mathbf{H}$ is the Hessian of f at $\mathbf{x}_k$.

- We can see that the local quadratic approximation contains the 1st-order Taylor approximation, while it is approximating the Hessian $\mathbf{H}$ by the matrix $\frac{1}{\alpha_k}\mathbf{I}$.

... continue

- Using $\frac{1}{\alpha_k}\mathbf{I}$ to approximate the Hessian $\mathbf{H}$ has the following physical meaning:
  - Assuming the function f is L-smooth: i.e., the gradient ∇f is L-Lipschitz
$$\|\nabla f(\mathbf{a}) - \nabla f(\mathbf{b})\|_2 \leq L\|\mathbf{a} - \mathbf{b}\|_2. \tag{2}$$
  - (2) means $\frac{\|\nabla f(\mathbf{a}) - \nabla f(\mathbf{b})\|_2}{\|\mathbf{a} - \mathbf{b}\|_2} \leq L$ for any $\mathbf{a}, \mathbf{b}$, hence it means the rate of change of ∇f is bounded, in other words, the Hessian of f is bounded.
  - Then, let $\alpha = \frac{1}{L}$, we have $\frac{1}{\alpha_k}\mathbf{I} = L\mathbf{I}$. If such L is in fact the largest eigenvalue of $\mathbf{H}$, then we have
$$L\mathbf{I} \succeq \mathbf{H},$$
    that is, $\frac{1}{\alpha_k}\mathbf{I} - \mathbf{H}$ is positive (semi)-definite.
  - Hence, when we set the stepsize α as the inverse of the Lipschitz constant L of the gradient ∇f, we are actually considering a local quadratic function that is the "tightest majorizer" of the 2nd-order Taylor approximation of f at $\mathbf{x}_k$ — the majorization-minimization view of GD.

A theorem

Theorem $\mathbf{x}_k - \alpha_k \nabla f(\mathbf{x}_k) = \underset{\mathbf{x}}{\operatorname{argmin}}\ F(\mathbf{x}; \mathbf{x}_k)$.

Proof. Direct proof, just solve for the minimizer of $F(\mathbf{x}; \mathbf{x}_k)$.

$$F(\mathbf{x}; \mathbf{x}_k) = f(\mathbf{x}_k) + (\mathbf{x} - \mathbf{x}_k)^\top \nabla f(\mathbf{x}_k) + \frac{1}{2\alpha_k} \|\mathbf{x} - \mathbf{x}_k\|_2^2.$$

Take derivative w.r.t. $\mathbf{x}$

$$\frac{\partial F(\mathbf{x}; \mathbf{x}_k)}{\partial \mathbf{x}} = \frac{\partial f(\mathbf{x}_k)}{\partial \mathbf{x}} + \frac{\partial}{\partial \mathbf{x}} (\mathbf{x} - \mathbf{x}_k)^\top \nabla f(\mathbf{x}_k) + \frac{\partial}{\partial \mathbf{x}} \frac{1}{2\alpha_k} \|\mathbf{x} - \mathbf{x}_k\|_2^2.$$

$\frac{\partial f(\mathbf{x}_k)}{\partial \mathbf{x}}$ = zero vector as $f(\mathbf{x}_k)$ is scalar, $\frac{\partial}{\partial \mathbf{x}} (\mathbf{x} - \mathbf{x}_k)^\top \nabla f(\mathbf{x}_k) = \nabla f(\mathbf{x}_k)$ and $\frac{\partial}{\partial \mathbf{x}} \frac{1}{2\alpha_k} \|\mathbf{x} - \mathbf{x}_k\|_2^2 = \frac{1}{\alpha_k} (\mathbf{x} - \mathbf{x}_k)$ so

$$\frac{\partial F(\mathbf{x}; \mathbf{x}_k)}{\partial \mathbf{x}} = \nabla f(\mathbf{x}_k) + \frac{1}{\alpha_k} (\mathbf{x} - \mathbf{x}_k).$$

Set the gradient to zero gives $\mathbf{x} = \mathbf{x}_k - \alpha_k \nabla f(\mathbf{x}_k)$. $\square$

Compact formulation ... (1/2)

- Consider $\mathbf{x}_{k+1} = \underset{\mathbf{x}}{\operatorname{argmin}}\ F(\mathbf{x}; \mathbf{x}_k)$, where

$$F(\mathbf{x}; \mathbf{x}_k) = f(\mathbf{x}_k) + (\mathbf{x} - \mathbf{x}_k)^\top \nabla f(\mathbf{x}_k) + \frac{1}{2\alpha_k} \|\mathbf{x} - \mathbf{x}_k\|_2^2.$$

- Expand

$$F(\mathbf{x}; \mathbf{x}_k) = \mathbf{x}^\top \nabla f(\mathbf{x}_k) + \frac{1}{2\alpha_k} \|\mathbf{x}\|_2^2 - \frac{1}{\alpha_k} \mathbf{x}^\top \mathbf{x}_k + \text{constants}.$$

- As adding or ignoring terms independent of $\mathbf{x}$ will not change the problem, so we have

$$\mathbf{x}_{k+1} = \underset{\mathbf{x}}{\operatorname{argmin}}\ \left\{ \mathbf{x}^\top \nabla f(\mathbf{x}_k) + \frac{1}{2\alpha_k} \|\mathbf{x}\|_2^2 - \frac{1}{\alpha_k} \mathbf{x}^\top \mathbf{x}_k \right\}.$$

Compact formulation ... (2/2)

- Factorize $\frac{1}{2\alpha_k}$ out, rearrange

$$\mathbf{x}_{k+1} = \underset{\mathbf{x}}{\operatorname{argmin}} \left\{ \frac{1}{2\alpha_k} \left(\|\mathbf{x}\|_2^2 - 2\mathbf{x}^\top \left(\mathbf{x}_k - \alpha_k \nabla f(\mathbf{x}_k) \right) \right) \right\}.$$

- Completing the squares and noting that adding or ignoring terms independent of $\mathbf{x}$ will not change the problem, we have

$$\mathbf{x}_{k+1} = \underset{\mathbf{x}}{\operatorname{argmin}} \frac{1}{2\alpha_k} \left\| \mathbf{x} - \left(\mathbf{x}_k - \alpha_k \nabla f(\mathbf{x}_k) \right) \right\|_2^2.$$

  which confirmed again $\mathbf{x} = \mathbf{x}_k - \alpha_k \nabla f(\mathbf{x}_k)$.

Application to regularized models

- The viewpoint of local quadratic model to understand GD can be used to derive update rule for model with regularization.

- Recall that for problem $\min_{\mathbf{x}} f(\mathbf{x})$, GD step can be written as

$$\mathbf{x}_{k+1} = \operatorname*{argmin}_{\mathbf{x}} \frac{1}{2\alpha_k} \left\| \mathbf{x} - \Big(\mathbf{x}_k - \alpha_k \nabla f(\mathbf{x}_k) \Big) \right\|_2^2$$

- For problem $\min_{\mathbf{x}} f(\mathbf{x}) + g(\mathbf{x})$, we can try

$$\mathbf{x}_{k+1} = \operatorname*{argmin}_{\mathbf{x}} \frac{1}{2\alpha_k} \left\| \mathbf{x} - \Big(\mathbf{x}_k - \alpha_k \nabla f(\mathbf{x}_k) \Big) \right\|_2^2 + g(\mathbf{x})$$

- We can see the addition of g alters the gradient update of $\mathbf{x}_k$ as

$$\frac{\partial}{\partial \mathbf{x}} \frac{1}{2\alpha_k} \left\| \mathbf{x} - \Big(\mathbf{x}_k - \alpha_k \nabla f(\mathbf{x}_k) \Big) \right\|_2^2 + \frac{\partial}{\partial \mathbf{x}} g(\mathbf{x}) = 0$$

  (if $g(\mathbf{x})$ is not differentiable, replace gradient by sub-gradient). This is useful for g that is non-smooth (not differentiable), e.g., L_1 regularized problems.

Last page - summary

- For $\min_{\mathbf{x}} f(\mathbf{x})$, the GD step
  $\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha_k \nabla f(\mathbf{x}_k) = \operatorname{argmin}_{\mathbf{x}} F(\mathbf{x}; \mathbf{x}_k)$
- $F(\mathbf{x}; \mathbf{x}_k)$ is a local quadratic approximation of of f at $\mathbf{x}_k$ as

$$F(\mathbf{x}; \mathbf{x}_k) = f(\mathbf{x}_k) + (\mathbf{x} - \mathbf{x}_k)^\top \nabla f(\mathbf{x}_k) + \frac{1}{2\alpha_k} \|\mathbf{x} - \mathbf{x}_k\|_2^2$$

- Compact form

$$\mathbf{x}_{k+1} = \operatorname*{argmin}_{\mathbf{x}} \left\{ \frac{1}{2\alpha_k} \left\| \mathbf{x} - \Big(\mathbf{x}_k - \alpha_k \nabla f(\mathbf{x}_k) \Big) \right\|_2^2 \right\}$$

- Application to problem $\min_{\mathbf{x} \in \mathbb{R}^n} f(\mathbf{x}) + g(\mathbf{x})$

$$\mathbf{x}_{k+1} = \operatorname*{argmin}_{\mathbf{x}} \left\{ \frac{1}{2\alpha_k} \left\| \mathbf{x} - \Big(\mathbf{x}_k - \alpha_k \nabla f(\mathbf{x}_k) \Big) \right\|_2^2 + g(\mathbf{x}) \right\},$$

  where update rule can be derived from the optimality condition.

End of document