Optimization for Deep Learning

Lecture 4-2: Proximal Gradient Descent

Kun Yuan

Peking University

Main contents in this lecture

- Proximal gradient descent

- Convergence properties

Optimization with simple regularizers

- Consider the following minimization problem with a regularizer

$$\min_{x \in \mathbb{R}^d} \quad f(x) + R(x) \tag{1}$$

- We assume $f(x)$ is L-smooth

- We assume regularizer $R(x)$ is closed, proper, and convex

- $R(x)$ can be non-differentiable, e.g., $R(x) = \|x\|_1$

Application: Robust principal component analysis

- Given an input matrix $M \in \mathbb{R}^{n \times d}$, we will find valuable information from M
- Consider the following problem[1]

$$\min_{L,S} \quad \frac{1}{2}\|M - (L+S)\|_F^2 + \lambda_1\|L\|_* + \lambda_2\|S\|_1$$

  - variable L represents **low-rank** background information; the nuclear-norm regularizer will promote its low-rank structure
  - variable S represents **sparse** valuable information; the ℓ_1-norm regularizer will promote its sparse structure
  - λ_1 and λ_2 are regularizer coefficients

[1] If we solve the problem with alternating minimization, then each subproblem is in the shape of problem (1).

Application: Robust principal component analysis

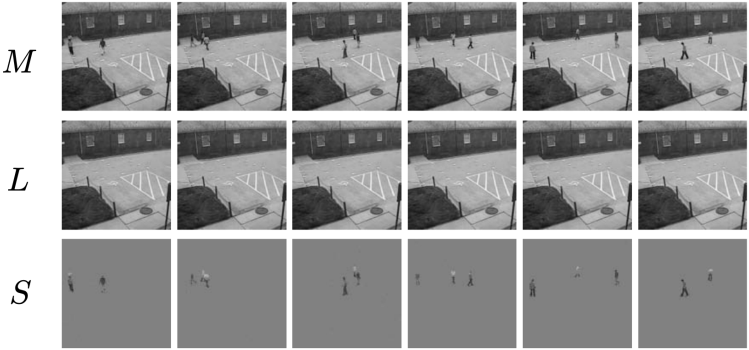

Figure: Split the input to low-rank and sparse components

Subgradient and subdifferential

Definition 1

Let $\psi : \mathbb{R}^d \to \mathbb{R}$ be a non-differentible function. It holds that $g \in \mathbb{R}^d$ is a ***subgradient*** *of ψ at x if and only if*

$$\psi(y) \geq \psi(x) + \langle g, y - x \rangle \quad \forall y \in \mathbb{R}^d$$

The set of subgradients of ψ at x is called the ***subdifferential*** *of x and is denoted by $\partial\psi(x)$.*

Examples:

- ℓ_1-norm: $\forall x \in \mathbb{R}^d$, $\psi(x) = \|x\|_1$, $\partial\psi(0) = \{g \in \mathbb{R}^d \mid |g_i| \leq 1, i = 1, \ldots, d\}$
- ℓ_2-norm: $\forall x \in \mathbb{R}^d$, $\psi(x) = \|x\|_2$, $\partial\psi(0) = \{g \in \mathbb{R}^d \mid \|g\|_2 \leq 1\}$.

Subgradient and subdifferential

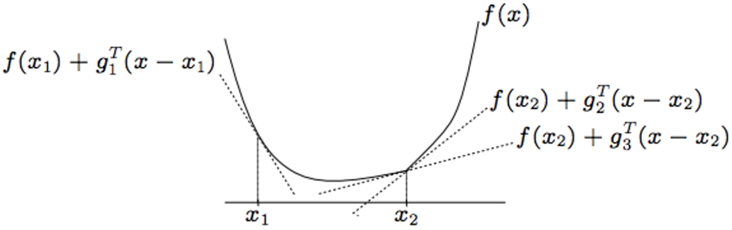

Figure: Illustration of the subgradient[2].

Subgradient reduces to gradient if ψ is differentiable at x

[2] Image is from wikipedia

Optimality conditions

Theorem 1

We suppose $\psi(x)$ is a convex and proper function. It holds that $x^\star$ is a global minimum of $\psi(x)$ if and only if

$$0 \in \partial\psi(x^\star).$$

Proximal gradient descent

- The main challenge is to handle the **non-differentible** regularizer

- We approximate $f(x)$ with a quadratic function:

$$f(x) \approx f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + \frac{1}{2\gamma} \|x - x_k\|^2$$

- Using the above approximation to replace $f(x)$, we have

$$\min_{x \in \mathbb{R}^d} \quad f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + \frac{1}{2\gamma} \|x - x_k\|^2 + R(x)$$

which is equivalent to

$$\min_{x \in \mathbb{R}^d} \quad R(x) + \frac{1}{2\gamma} \left\|x - \big(x_k - \gamma \nabla f(x_k)\big)\right\|^2$$

Proximal gradient descent

- Continue the procedure, we achieve proximal gradient descent

$$y_{k+1} = x_k - \gamma \nabla f(x_k)$$
$$x_{k+1} = \mathrm{prox}_{\gamma R}(y_{k+1})$$

  where the **proximity operator** $\mathrm{prox}_h(\cdot)$ is defined as

$$\mathrm{prox}_h(x) := \arg\min_{u \in \mathbb{R}^d} \{h(x) + \frac{1}{2}\|x - u\|^2\}$$

- Throughout the lecture, we assume $R(x)$ is an easy regularizer, i.e., the proximity operator has a **closed-form** solution.

Can proximal GD converge to the solution? Yes!

Lemma 1

Suppose $R(x)$ is proper closed and convex. If proximal gradient descent converges to a fixed point, i.e.,

$$x^\star = \mathrm{prox}_{\gamma R}(x^\star - \gamma \nabla f(x^\star)),$$

then it holds that

$$0 \in \nabla f(x^\star) + \partial R(x^\star)$$

Proof: $x^\star = \mathrm{prox}_{\gamma R}(x^\star - \gamma \nabla f(x^\star)) \Longleftrightarrow 0 \in \gamma \partial R(x^\star) + x^\star - (x^\star - \gamma \nabla f(x^\star))$

If $f(x)$ is convex, the fixed point is the global minimum.

Examples of easy regularizers

- ℓ_1-norm: $\forall x \in \mathbb{R}^d$, $R(x) = \|x\|_1$, $[\mathrm{prox}_{\gamma R}(x)]_i = \mathsf{sign}(x_i) \max\{|x_i| - \gamma, 0\}$.

- ℓ_2-norm: $\forall x \in \mathbb{R}^d$, $R(x) = \|x\|_2$,

$$\mathrm{prox}_{\gamma R}(x) = \begin{cases} \left(1 - \frac{\gamma}{\|x\|_2}\right)x, & \|x\|_2 \geq R, \\ 0, & \text{otherwise.} \end{cases}$$

- Projection: Let $\mathcal{C}$ be a closed convex set and $I_{\mathcal{C}}(x)$ is an indicator function

$$\begin{aligned} \mathrm{prox}_{I_{\mathcal{C}}}(x) &= \arg\min_u \left\{ I_{\mathcal{C}}(u) + \frac{1}{2}\|u - x\|^2 \right\} \\ &= \arg\min_{u \in \mathcal{C}} \|u - x\|^2 \\ &= \mathcal{P}_{\mathcal{C}}(x) \end{aligned}$$

Projected GD is a special example of proximal GD

- Recall the constraind minimization problem

$$\min_{x\in\mathbb{R}^d} \quad f(x) \quad \text{subject to} \quad x \in \mathcal{X}$$

  where $\mathcal{X}$ is a closed convex set.

- With indicator function, we can reformulate it as

$$\min_{x\in\mathbb{R}^d} \quad f(x) + I_{\mathcal{X}}(x)$$

- Projected GD is essentially proximal GD

$$\begin{aligned} x_{k+1} &= \mathrm{prox}_{I_{\mathcal{X}}}[x_k - \gamma \nabla f(x_k)] \quad && \text{(Proximal GD)} \\ &= \mathcal{P}_{\mathcal{X}}[x_k - \gamma \nabla f(x_k)] && \text{(Projected GD)} \end{aligned}$$

Convergence: Smooth and strongly-convex scenario

Lemma 2

If $R(x)$ is a closed convex proper function, then

$$\|\text{prox}_R(x) - \text{prox}_R(y)\| \leq \|x - y\|.$$

It implies that $\text{prox}_R(x)$ is non-expansive.

We leave it as an exercise.

Convergence: Smooth and strongly-convex scenario

Lemma 3

If $f(x)$ is convex and differentiable, and $R(x)$ is a closed convex proper function, then the optimal solution $x^\star = \arg\min_{x \in \mathbb{R}^d} \{f(x) + R(x)\}$ satisfies

$$x^\star = \mathrm{prox}_{\gamma R}(x^\star - \gamma \nabla f(x^\star))$$

Easy to show. We leave it as an exercise.

Convergence: Smooth and strongly-convex scenario

Theorem 2

We assume $f(x)$ is μ-strongly convex and L-smooth on $\mathbb{R}^d$, $R(x)$ is a closed convex proper function, and $x^\star$ is the optimal solution. If we set $\gamma = 1/L$, proximal gradient descent with an arbitrary x_0 satisfies

$$\|x^K - x^\star\| \le (1 - \frac{\mu}{L})^K \|x^0 - x^\star\|.$$

Easy to show. We leave it as an exercise.

Projected GD has a rate $O((1-\mu/L)^K)$ and a complexity $O(L/\mu \log(1/\epsilon))$

It has the same order in rate and complexity as gradient descent

Convergence: Smooth and convex scenario

We let $\psi(x) = f(x) + R(x)$ and $\psi^\star = \psi(x^\star)$

Theorem 3

We assume $f(x)$ is L-smooth on $\mathbb{R}^d$, $R(x)$ is a closed convex proper function. If we set $\gamma = 1/L$, proximal gradient descent with an arbitrary x_0 satisfies

$$\psi(x^K) - \psi^\star \leq \frac{L}{2K}\|x_0 - x^\star\|^2.$$

Projected GD has a rate $O(L/K)$, which amounts to complexity $O(L/\epsilon)$

It has the same order in rate and complexity as gradient descent

Comparison between GD and proximal GD

Method	Convexity	Rate	Complexity
GD	Non-convex	$O(L/k)$	$O(L/\epsilon)$
	Convex	$O(L/k)$	$O(L/\epsilon)$
	Strongly convex	$O((1-\frac{\mu}{L})^k)$	$O(\frac{L}{\mu}\log(1/\epsilon))$
Proximal GD	Non-convex	$O(L/k)$	$O(L/\epsilon)$
	Convex	$O(L/k)$	$O(L/\epsilon)$
	Strongly convex	$O((1-\frac{\mu}{L})^k)$	$O(\frac{L}{\mu}\log(1/\epsilon))$

Proximal GD converges as fast as GD even with the projection step. It makes sense since both GD and projected GD are special examples of proximal GD.

Summary

- Optimizaiton with simple regularizers are common in applications
- Proximal GD is very useful when $\mathrm{prox}_R(\cdot)$ is cheap.
- Proximal GD has the same convergence rate and complexity as GD.