Stochastic Subgradient Method

Dr. Dingzhu Wen

School of Information Science and Technology (SIST)
ShanghaiTech University

wendzh@shanghaitech.edu.cn

March 20, 2024

Overview

Review of Subgradient Method

Optimization Problem: $\min_{\mathbf{w} \in \mathbb{R}^N} \ \mathcal{L}(\mathbf{w})$.

- $\mathcal{L}(\mathbf{w})$ is convex and non-differentiable.
- Definition of subgradient: A subgradient of a function $\mathcal{L} : \mathbb{R}^N \to \mathbb{R}$ at $\mathbf{w}_1$ is any vector that satisfies

$$\mathcal{L}(\mathbf{w}_2) \geq \mathcal{L}(\mathbf{w}_1) + \mathbf{g}^T(\mathbf{w}_2 - \mathbf{w}_1), \quad \forall \mathbf{w}_2.$$

- Subgradient method: $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta \mathbf{g}_t$.
- Subgradient method is not a descent method: the function value can (and often does) increase.

$$\mathcal{L}(\mathbf{w}_{\mathrm{best},T}) = \min_{t=0,1,\ldots,T} \mathcal{L}(\mathbf{w}_t).$$

Review of Subgradient Method

Updating Settings

- Fixed step size: $\eta_t = \eta$.
- Diminishing step size: $\sum_{t=1}^{+\infty} \eta_t^2 < +\infty$, $\sum_{t=1}^{+\infty} \eta_t = +\infty$.

Assumption

- $\mathcal{L}(\cdot)$ is convex.
- $\mathcal{L}(\cdot)$ is Lipschitz contimuous:

$$\mathcal{L}(\mathbf{w}_2) - \mathcal{L}(\mathbf{w}_1) \leq G \left\| \mathbf{w}_2 - \mathbf{w}_1 \right\| .$$

Review of Subgradient Method

Convergence

Theorem

For a fixed step size η, subgradient method $\{\mathbf{w}_{t+1} = \mathbf{w}_t - \eta \mathbf{g}_t\}$ satisfies

$$\lim_{T \to +\infty} \mathcal{L}(\mathbf{w}_{\mathrm{best},T}) \leq \mathcal{L}_* + \frac{G^2 \eta}{2}.$$

Theorem

For a diminishing step size, subgradient method $\{\mathbf{w}_{t+1} = \mathbf{w}_t - \eta \mathbf{g}_t\}$ satisfies

$$\lim_{T \to +\infty} \mathcal{L}(\mathbf{w}_{\mathrm{best},T}) = \mathcal{L}_*.$$

Proximal Gradient Method

Learning Loss: $\mathcal{L}(\mathbf{w}) = f(\mathbf{w}) + r(\mathbf{w})$,

- $f(\mathbf{w})$, objective function (learning model), e.g., SVM, logistic regression,
- $r(\mathbf{w})$, regularization function.

Regularization:

- ℓ_1 regularization: $r(\mathbf{w}) = \lambda \|\mathbf{w}\|_1$,
  - ℓ_1 norm, $\|\mathbf{w}\|_1 = \sum_i |w_i|$,
  - Advatange: Avoid overfitting, enhance sparsity,
  - Shortage: Non-differentiable, low convergence rate.
- ℓ_2 regularization: $r(\mathbf{w}) = \lambda \|\mathbf{w}\|_2$,
  - Advatange: Avoid overfitting, Differentiable,
  - Shortage: Sparsity not guaranteed.

Proximal Gradient Method

ℓ_1-regularized Loss: $\mathcal{L}(\mathbf{w}) = f(\mathbf{w}) + r(\mathbf{w})$,

- $f(\mathbf{w})$ and $r(\mathbf{w})$ are convex.
- $r(\mathbf{w})$ is non-differentiable.

Question: How to find a good subgradient method?

- How to find a sub-gradient? Not derivatives.
- which is a good sub-gradient, leading to fast convergence?

Answer: Proximal method.

Proximal Gradient Method

If $\mathcal{L}(\mathbf{w})$ is differentiable, the following two updating ways are equivalent:

- GD: $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta \nabla \mathcal{L}(\mathbf{w}_t)$,
- $\mathbf{w}_{t+1} = \arg\min_{\mathbf{z}} \; \mathcal{L}(\mathbf{w}_t) + \nabla \mathcal{L}(\mathbf{w}_t)^T (\mathbf{z} - \mathbf{w}_t) + \frac{1}{2\eta} \|\mathbf{z} - \mathbf{w}_t\|^2$.

If $\mathcal{L}(\mathbf{w}) = f(\mathbf{w}) + r(\mathbf{w})$ is non-differentiable, **proximal gradient method**:

$$\mathbf{w}_{t+1} = \arg\min_{\mathbf{z}} \; f(\mathbf{w}_t) + \nabla f(\mathbf{w}_t)^T (\mathbf{z} - \mathbf{w}_t) + \frac{1}{2\eta} \|\mathbf{z} - \mathbf{w}_t\|^2 + r(\mathbf{z}).$$

Proximal gradient method has the property of descent iterations and outperforms the general sub-gradient methods.

Proximal Gradient Method

Proximal Mapping

$$\mathrm{Prox}_{r,\eta}(\mathbf{x}) = \arg\min_{\mathbf{z}} \ \frac{1}{2\eta}\|\mathbf{z}-\mathbf{x}\|_2^2 + r(\mathbf{z}).$$

Proximal Gradient:

$$\begin{aligned}\mathbf{w}_{t+1} &= \mathrm{Prox}_{r,\eta}(\mathbf{w}_t - \eta\nabla f(\mathbf{w}_t)),\\ &= \arg\min_{\mathbf{z}} \ \frac{1}{2\eta}\|\mathbf{z}-(\mathbf{w}_t - \eta\nabla f(\mathbf{w}_t))\|_2^2 + r(\mathbf{z}).\end{aligned}$$

- The proximal map $\mathrm{Prox}_{r,\eta}(\cdot)$ can be computed analytically for a lot of r functions.
- $\mathrm{Prox}_{r,\eta}(\cdot)$ does not depend on the leraning objective $f(\cdot)$, only on r.
- $f(\cdot)$ can be a complicated function, all we need to do is to compute its gradient.

Proximal Mapping

Proximal Mapping: $\mathrm{Prox}_{r,\eta}(\mathbf{x}) = \arg\min_{\mathbf{z}} \ \frac{1}{2\eta}\|\mathbf{z}-\mathbf{x}\|_2^2 + r(\mathbf{z})$,

- $r(\cdot)$ is convex and closed.

Theorem (Existence and Uniquess)

$\mathrm{Prox}_{r,\eta}(\mathbf{x})$ *exsits and is unique for all* $\mathbf{x}$*.*

Proof: $r(\mathbf{z})$ is closed and convex (Subgradient equals to 0 only in one point).

Theorem (Proximal Gradient is Subgradient)

$\frac{1}{\eta}[\mathbf{x} - \mathrm{Prox}_{r,\eta}(\mathbf{x})]$ *is a subgradient of* $r(\mathbf{z})$*.*

Proof: $0 \in \partial\left[\frac{1}{2\eta}\|\mathbf{z}-\mathbf{x}\|_2^2 + r(\mathbf{z})\right]$. This indicates proximal method belongs to subgradient method.

Iterative soft-thresholding algorithm (ISTA)

Proximal Mapping

$$\begin{aligned}\mathrm{Prox}_{r,\eta}(\mathbf{x}) &= \arg\min_{\mathbf{z}} \ \frac{1}{2\eta}\|\mathbf{z}-\mathbf{x}\|_2^2 + \lambda\|\mathbf{z}\|_1 ,\\ &= \arg\min_{\mathbf{z}} \ \frac{1}{2}\|\mathbf{z}-\mathbf{x}\|_2^2 + \eta\lambda\|\mathbf{z}\|_1 ,\\ &\triangleq \mathbf{s}_{\lambda,\eta}(\mathbf{x}).\end{aligned}$$

The i-th element of $\mathbf{s}_{\lambda,\eta}(\mathbf{x})$ is a **soft-thresholding operator**:

$$s_i(x_i) = \begin{cases} x_i - \lambda\eta, & \text{if } x_i > \lambda\eta, \\ 0, & \text{if } -\lambda\eta \le x_i \le \lambda\eta, \\ x_i + \lambda\eta, & \text{if } x_i < -\lambda\eta. \end{cases}$$

Example: Lasso

Lasso Problem: $\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{y}-\mathbf{X}\mathbf{w}\|_2^2+\lambda\|\mathbf{w}\|_1$.

- $f(\mathbf{w})=\frac{1}{2}\|\mathbf{y}-\mathbf{X}\mathbf{w}\|_2^2$,
- $r(\mathbf{w})=\lambda\|\mathbf{w}\|_1$,
- $\mathbf{y}$ is the label vector,
- $\mathbf{X}$ is the collection of feature vector.

Gradient of $f(\mathbf{w})$: $\nabla f(\mathbf{w})=-\mathbf{X}^T(\mathbf{y}-\mathbf{X}\mathbf{w})$.
Proximal Gradient Update:

$$\mathbf{w}_{t+1}=\mathbf{s}_{\lambda,\eta}\left(\mathbf{w}_t+\mathbf{X}^T(\mathbf{y}-\mathbf{X}\mathbf{w})\right).$$

Convergence Analysis

Assumptions

Assumption

(A1: L-Smoothness) $f(\mathbf{w})$ is L-smooth:

$$\left\|\nabla f(\mathbf{w}_1) - \nabla f(\mathbf{w}_2)\right\|_2 \leq L \left\|\mathbf{w}_1 - \mathbf{w}_2\right\|_2, \quad \forall(\mathbf{w}_1, \mathbf{w}_2).$$

(A2:Lower Bounded) $f(\mathbf{w})$ is bounded: The optimum f_* is finite and is attained at $\mathbf{w}_*$ (Not necessarily unique).

Affine Lower Bound from Convexity:

$$f(\mathbf{w}_2) \geq f(\mathbf{w}_1) + \nabla f(\mathbf{w}_1)^T(\mathbf{w}_2 - \mathbf{w}_1),\ \forall(\mathbf{w}_1, \mathbf{w}_2).$$

Quadratic Upper Bound from L-Smoothness:

$$f(\mathbf{w}_2) \leq f(\mathbf{w}_1) + \nabla f(\mathbf{w}_1)^T(\mathbf{w}_2 - \mathbf{w}_1) + \frac{L}{2}\left\|\mathbf{w}_1 - \mathbf{w}_2\right\|_2^2,\ \forall(\mathbf{w}_1, \mathbf{w}_2).$$

Convergence Analysis

Denote $\mathbf{G}_\eta(\mathbf{w}) = \frac{1}{\eta}\left[\mathbf{w} - \mathrm{Prox}_{r,\eta}(\mathbf{w} - \eta\nabla f(\mathbf{w}))\right]$.

Proximal Gradient Iteration:

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta\mathbf{G}_\eta(\mathbf{w}_t).$$

At the optimum, $\mathbf{G}_\eta(\mathbf{w}_*) = \mathbf{0}$.

- According to the L-smoothness of $f(\cdot)$,

$$f(\mathbf{w}_{t+1}) \le f(\mathbf{w}_t) + \nabla f(\mathbf{w}_t)^T(\mathbf{w}_{t+1} - \mathbf{w}_t) + \frac{L}{2}\left\|\mathbf{w}_{t+1} - \mathbf{w}_t\right\|_2^2.$$

- It follows that

$$f(\mathbf{w}_{t+1}) \le f(\mathbf{w}_t) - \eta\nabla f(\mathbf{w}_t)^T\mathbf{G}_\eta(\mathbf{w}_t) + \frac{L\eta^2}{2}\left\|\mathbf{G}_\eta(\mathbf{w}_t)\right\|_2^2.$$

Convergence Analysis

- For step size $0 \leq \eta \leq 1/L$,

$$f(\mathbf{w}_{t+1}) \leq f(\mathbf{w}_t) - \eta \nabla f(\mathbf{w}_t)^T \mathbf{G}_\eta(\mathbf{w}_t) + \frac{\eta}{2} \|\mathbf{G}_\eta(\mathbf{w}_t)\|_2^2 .$$

- Then, for all $\mathbf{z}$,

$$\mathcal{L}(\mathbf{w}_{t+1}) \leq \mathcal{L}(\mathbf{z}) + \mathbf{G}_\eta(\mathbf{w}_t)^T(\mathbf{w}_t - \mathbf{z}) - \frac{\eta}{2} \|\mathbf{G}_\eta(\mathbf{w}_t)\|_2^2 .$$

Proof (With $\mathbf{v} = \mathbf{G}_\eta(\mathbf{w}_t) - \nabla f(\mathbf{w}_t)$):

$$\begin{aligned}
\mathcal{L}(\mathbf{w}_{t+1}) &= f(\mathbf{w}_{t+1}) + r(\mathbf{w}_{t+1}), \\
&\leq f(\mathbf{w}_t) - \eta \nabla f(\mathbf{w}_t)^T \mathbf{G}_\eta(\mathbf{w}_t) + \frac{\eta}{2} \|\mathbf{G}_\eta(\mathbf{w}_t)\|_2^2 + r(\mathbf{w}_{t+1}),
\end{aligned}$$

Convergence Analysis

- Then, for all $\mathbf{z}$,

$$\mathcal{L}(\mathbf{w}_{t+1}) \leq \mathcal{L}(\mathbf{z}) + \mathbf{G}_\eta(\mathbf{w}_t)^T(\mathbf{w}_t - \mathbf{z}) - \frac{\eta}{2}\|\mathbf{G}_\eta(\mathbf{w}_t)\|_2^2.$$

Proof (With $\mathbf{v} = \mathbf{G}_\eta(\mathbf{w}_t) - \nabla f(\mathbf{w}_t)$, **Continue**):

$$\begin{aligned}\mathcal{L}(\mathbf{w}_{t+1}) \leq& f(\mathbf{z}) + \nabla f(\mathbf{w}_t)^T(\mathbf{w}_t - \mathbf{z}) - \eta\nabla f(\mathbf{w}_t)^T\mathbf{G}_\eta(\mathbf{w}_t) \\ &+ \frac{\eta}{2}\|\mathbf{G}_\eta(\mathbf{w}_t)\|_2^2 + r(\mathbf{z}) + \mathbf{v}^T(\mathbf{w}_t - \mathbf{z} - \eta\mathbf{G}_\eta(\mathbf{w}_t)), \\ =& f(\mathbf{z}) + r(\mathbf{z}) + \mathbf{G}_\eta(\mathbf{w}_t)^T(\mathbf{w}_t - \mathbf{z}) - \frac{\eta}{2}\|\mathbf{G}_\eta(\mathbf{w}_t)\|_2^2.\end{aligned}$$

Line 2 holds due to convexity of $f(\cdot)$ and $r(\cdot)$. Besides, $\mathbf{v} \in \partial r(\mathbf{w}_t - \eta\mathbf{G}_\eta(\mathbf{w}_t))$.

Convergence Analysis

- By taking $\mathbf{z} = \mathbf{w}_t$,

$$\mathcal{L}(\mathbf{w}_{t+1}) \leq \mathcal{L}(\mathbf{w}_t) - \frac{\eta}{2} \|\mathbf{G}_\eta(\mathbf{w}_t)\|_2^2 ,$$

  which means that the proximal gradient method is actually a descent method. (However, general sub-gradient methods are not guaranteed to be descent.)

- By taking $\mathbf{z} = \mathbf{w}_*$,

$$\begin{aligned}\mathcal{L}(\mathbf{w}_{t+1}) - \mathcal{L}_* &\leq \mathbf{G}_\eta(\mathbf{w}_t)^T(\mathbf{w}_t - \mathbf{w}_*) - \frac{\eta}{2} \|\mathbf{G}_\eta(\mathbf{w}_t)\|_2^2 ,\\ &= \frac{1}{2\eta}\left[\|\mathbf{w}_t - \mathbf{w}_*\|^2 - \|\mathbf{w}_t - \mathbf{w}_* - \eta\mathbf{G}_\eta(\mathbf{w}_t)\|^2\right],\\ &= \frac{1}{2\eta}\left[\|\mathbf{w}_t - \mathbf{w}_*\|^2 - \|\mathbf{w}_{t+1} - \mathbf{w}_*\|^2\right]\end{aligned}$$

  It follows that $|\mathbf{w}_{t+1} - \mathbf{w}_*\|^2 \leq \|\mathbf{w}_t - \mathbf{w}_*\|^2$.

Convergence Analysis

Analysis for Fixed Step Size

$$
\begin{aligned}
\sum_{t=0}^{T-1} \mathcal{L}(\mathbf{w}_{t+1}) - \mathcal{L}_* &\leq \sum_{t=0}^{T-1} \frac{1}{2\eta} \left[\|\mathbf{w}_t - \mathbf{w}_*\|^2 - \|\mathbf{w}_{t+1} - \mathbf{w}_*\|^2 \right], \\
&= \frac{1}{2\eta} \left[\|\mathbf{w}_0 - \mathbf{w}_*\|^2 - \|\mathbf{w}_T - \mathbf{w}_*\|^2 \right], \\
&\leq \frac{1}{2\eta} \|\mathbf{w}_0 - \mathbf{w}_*\|^2 .
\end{aligned}
$$

As it's a descent method,

$$
\mathcal{L}(\mathbf{w}_T) - \mathcal{L}_* \leq \frac{1}{2\eta T} \|\mathbf{w}_0 - \mathbf{w}_*\|^2 .
$$

Conclusion: Reaches $\mathcal{L}(\mathbf{w}_T) - \mathcal{L}_* \leq \epsilon$ after $\mathcal{O}(1/\epsilon)$ iterations.

Convergence Analysis

Analysis with Line Search: Adpative step sizes $\eta_t > \eta_{\min}$

$$\begin{aligned}\sum_{t=0}^{T-1} \mathcal{L}(\mathbf{w}_{t+1}) - \mathcal{L}_* &\leq \sum_{t=0}^{T-1} \frac{1}{2\eta_t} \left[\|\mathbf{w}_t - \mathbf{w}_*\|^2 - \|\mathbf{w}_{t+1} - \mathbf{w}_*\|^2 \right], \\ &= \frac{1}{2\eta_{\min}} \left[\|\mathbf{w}_0 - \mathbf{w}_*\|^2 - \|\mathbf{w}_T - \mathbf{w}_*\|^2 \right], \\ &\leq \frac{1}{2\eta_{\min}} \|\mathbf{w}_0 - \mathbf{w}_*\|^2 .\end{aligned}$$

As it's a descent method,

$$\mathcal{L}(\mathbf{w}_T) - \mathcal{L}_* \leq \frac{1}{2\eta_{\min} T} \|\mathbf{w}_0 - \mathbf{w}_*\|^2 .$$

Conclusion: Reaches $\mathcal{L}(\mathbf{w}_T) - \mathcal{L}_* \leq \epsilon$ after $\mathcal{O}(1/\epsilon)$ iterations.

Stochastic Subgradient Method

Learning Loss: $\mathcal{L}(\mathbf{w}) = \frac{1}{N}\sum_{i=1}^{N} f(\mathbf{w}; \mathbf{z}_i) + r(\mathbf{w})$,

- $f(\mathbf{w})$, objective function (learning model), e.g., SVM, logistic regression,
- $\mathbf{z}_i$, data sample,
- $r(\mathbf{w})$, non-differentiable regularization function.

Stochastic Subgradient Method:

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta \mathbf{g}_{i_t},$$

- $\mathbf{g}_{i_t} \in \partial f(\mathbf{w}; \mathbf{z}_{i_t})$,
- i_t is the index of the selected sample in the t-th iteration.

Convergence Analysis

Assumptions

Assumption

(A1: μ-Strongly Convex) $f(\mathbf{w})$ is μ-strongly congex.

(A2: Unbiased Estimation) $\mathbb{E}\left[\mathbf{g}_{i_t}|\mathbf{w}_t\right] = \mathbf{g}_t$.

(A3: Bounded Subgradient Norm) $\mathbb{E}\left[\|\mathbf{g}_{i_t}\|^2\right] \leq B^2$ (Finite variance and bounded subgradients).

Convergence Analysis

Expansion of distance:

$$\begin{aligned}\left\|\mathbf{w}_{t+1}-\mathbf{w}_*\right\|^2 &= \left\|\mathbf{w}_t-\eta\mathbf{g}_{i_t}-\mathbf{w}_*\right\|^2, \\ &= \left\|\mathbf{w}_t-\mathbf{w}_*\right\|^2-2\eta\mathbf{g}_{i_t}^T(\mathbf{w}_t-\mathbf{w}_*)+\eta^2\left\|\mathbf{g}_{i_t}\right\|^2.\end{aligned}$$

Take expectation:

$$\begin{aligned}\mathbb{E}\left[\left\|\mathbf{w}_{t+1}-\mathbf{w}_*\right\|^2\right] &= \left\|\mathbf{w}_t-\mathbf{w}_*\right\|^2-2\eta\mathbb{E}\left[\mathbf{g}_{i_t}^T\right](\mathbf{w}_t-\mathbf{w}_*)+\eta^2\mathbb{E}\left[\left\|\mathbf{g}_{i_t}\right\|^2\right], \\ &\leq \left\|\mathbf{w}_t-\mathbf{w}_*\right\|^2-2\eta\mathbf{g}_t^T(\mathbf{w}_t-\mathbf{w}_*)+\eta^2B^2\end{aligned}$$

μ-strongly convex: $(\mathbf{g}_t-\mathbf{0})^T(\mathbf{w}_t-\mathbf{w}_*) \geq \mu\left\|\mathbf{w}_t-\mathbf{w}_*\right\|^2$.

Convergence Analysis

It follows that

$$
\begin{aligned}
\mathbb{E}\left[\left\|\mathbf{w}_{t+1}-\mathbf{w}_*\right\|^2\right] \leq & \left\|\mathbf{w}_t-\mathbf{w}_*\right\|^2-2\eta\mu\left\|\mathbf{w}_t-\mathbf{w}_*\right\|^2+\eta^2 B^2, \\
= & (1-2\eta\mu)\left\|\mathbf{w}_t-\mathbf{w}_*\right\|^2+\eta^2 B^2.
\end{aligned}
$$

- Similar to linear convergence,
- Fixed step size leads to non-zero distance,
- Diminishing step size sacrifies the rate.

Thank you!

wendzh@shanghaitech.edu.cn