COMM 616: Modern Optimization with Applications in ML and OR 2024-25 Fall

Lecture 8: Subgradient Methods Under Weakly Convexity

Instructor: Jiajin Li Scribe: Zhuyu Liu

1 Summary of (Sub)gradient Methods under Different Conditions

Function Classes	Residual Function R	Covergence Rate	Step Size
Convex + L-Smooth	$f(\boldsymbol{x}^k) - f^*$	$O\left(\frac{1}{k}\right)$	Constant $1/L$
μ-Convex + L-Smooth	$f(\boldsymbol{x}^k) - f^*,\ \ \|\boldsymbol{x}^k - \boldsymbol{x}^*\|^2$	$O\left(\exp\left(-\frac{k}{\kappa}\right)\right)$	Constant $1/L$
μ-PL + L-Smooth	$f(\boldsymbol{x}^k) - f^*$	$O\left(\exp\left(-\frac{k}{\kappa}\right)\right)$	Constant $1/L$
L-Smooth	$\min_{k\in[K]} \|\nabla f(\boldsymbol{x}^k)\|$	$O\left(\frac{1}{\sqrt{k}}\right)$	Constant $1/L$
Convex + L-Lip	$\min_{k\in[K]} f(\boldsymbol{x}^k) - f^*$	$O\left(\frac{\log k}{\sqrt{k}}\right)$	$O\left(\frac{1}{\sqrt{k}}\right)$
Strongly Convex + L-Lip	$\min_{k\in[K]} f(\boldsymbol{x}^k) - f^*$	$O\left(\frac{1}{k}\right)$	$O\left(\frac{1}{k}\right)$
Weakly Convex	$\mathbb{E}\left[\|\nabla f_{\frac{1}{\hat{\rho}}}(\boldsymbol{x}^k)\|^2\right]$	$O\left(\frac{1}{k^{1/4}}\right)$	$O\left(\frac{1}{\sqrt{k}}\right)$

Table 1: Summary of (Sub)gradient Methods under Different Conditions

Remark 1. *Let κ be a constant number defined as L/μ.*

Remark 2. *When the problem is smooth, we can see that it is straightforward to extend from convex to nonconvex cases:*

$$f(\boldsymbol{x}^k) - f^*,\ \|\boldsymbol{x}^k - \boldsymbol{x}^*\| \quad \rightarrow \quad \|\nabla f(\boldsymbol{x}^k)\|$$

However, obtaining a valid stationary measure for nonsmooth cases is more challenging. Today, we will first discuss a subclass of nonsmooth nonconvex functions—weakly convex functions—and explore their stationary measures using the Moreau envelope technique.

2 Weakly Convexity and Modreau Envelope [1, 2]

Definition 3 (Weakly Convexity)**.** *A function $f : \mathbb{R}^d \to \bar{\mathbb{R}}$ is ρ-weakly convex if there exists $\rho > 0$ such that $f(\boldsymbol{x}) + \frac{\rho}{2}\|\boldsymbol{x}\|^2$ is convex, that is:*

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^\top(\boldsymbol{y} - \boldsymbol{x}) - \frac{\rho}{2}\|\boldsymbol{y} - \boldsymbol{x}\|^2 \quad \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d.$$

Fact 4. *L-Smooth functions are L-weakly convex functions.*

This categorizes L-smooth functions within the broader class of weakly convex functions, which includes nonsmooth ones. To extend the gradient norm to the nonsmooth case, one might consider the distance from zero to the subdifferential as a natural extension. However, this approach can lead to discontinuities. For example, with $f(\boldsymbol{x}) = |\boldsymbol{x}|$, we find that $\mathrm{dist}(0, \partial f(\boldsymbol{x})) = 1$ everywhere except at $\boldsymbol{x} = 0$. In other words, we cannot use this measure to quantify the near-stationarity. Luckily, we have the following observation.

Observation 5. *Weakly convex problems naturally admit a continuous measure of stationarity through implicit smoothing.*

Definition 6 (Moreau Envelope (Inf Projection))**.** *Suppose that $0 < \lambda < \rho^{-1}$. We define the Moreau envelope of f with respect to the modulus λ as follows:*

$$f_\lambda(\boldsymbol{x}) = \inf_{y} \left\{ f(\boldsymbol{y}) + \frac{1}{2\lambda}\|\boldsymbol{x} - \boldsymbol{y}\|^2 \right\}.$$

Remark 7. *As long as f is ρ-weakly convex and $\lambda < \rho^{-1}$, then the function $f_\lambda(\boldsymbol{x})$ is C^1-smooth with gradient:*

$$\nabla f_\lambda(\boldsymbol{x}) = \lambda^{-1}(x - \operatorname{prox}_{\lambda f}(\boldsymbol{x})),$$

where prox$(\cdot)_{\lambda f}$ is the proximal operator defined as

Definition 8 (Proximal Operator)**.** *The proximal operator of f at the point $\boldsymbol{x}$ with respect to the modulus $0 < \lambda < \rho^{-1}$ is defined as:*

$$\operatorname{prox}_{\lambda f}(\boldsymbol{x}) = \arg\min_{\boldsymbol{y}} \left\{ f(\boldsymbol{y}) + \frac{1}{2\lambda}\|\boldsymbol{y} - \boldsymbol{x}\|^2 \right\}$$

Theorem 9 (Envelope Theorem (Informal Version))**.** *Consider a $f : \mathbb{R}^d \to \mathbb{R}$ defined as:*

$$f(\boldsymbol{x}) = \max_{y \in \mathcal{Y}} g(\boldsymbol{x}, \boldsymbol{y}).$$

If $g(\cdot, \boldsymbol{y})$ is smooth for every $\boldsymbol{y} \in \mathcal{Y}$ and has a unique maximizer $y^\star(\boldsymbol{x})$ for each $\boldsymbol{x}$, then we know the function f is differentiable whose gradient can be computed as

$$\nabla f(\boldsymbol{x}) = \nabla_{\boldsymbol{x}} g(\boldsymbol{x}, \boldsymbol{y})\Big|_{\boldsymbol{y}=\boldsymbol{y}^\star(\boldsymbol{x})}.$$

Please also see the generalized version in [3, Theorem 10.31] (Danskin's Theorem).

Example 10. *Consider the following structured nonsmooth minimization problem:*

$$\min_{\boldsymbol{x}} \left[\max_{k \in [K]} f_k(\boldsymbol{x}) \right]$$

If each f_k is L-smooth, then the minimization objective is also weekly convex but non-smooth.

Question: Does $\|\nabla f_\lambda(\boldsymbol{x})\|$ have an intuitive interpretation in terms of the near-stationarity of the target problem $\inf_{\boldsymbol{x} \in \mathbb{R}^d} f(\boldsymbol{x})$?

Proposition 11. *For any $\boldsymbol{x} \in \mathbb{R}^d$, we define the proximal point $\hat{\boldsymbol{x}} :=$ prox$_{\lambda f}(\boldsymbol{x})$. Then, we have:*

(i) $\|\hat{\boldsymbol{x}} - \boldsymbol{x}\| = \lambda\|\nabla f_\lambda(\boldsymbol{x})\|$

(ii) $f(\hat{\boldsymbol{x}}) \le f(\boldsymbol{x})$

(iii) $\operatorname{dist}(0, \partial f(\boldsymbol{x})) \le \|\nabla f_\lambda(\boldsymbol{x})\|$.

Proof. The first statement can be easily proved by the fact:

$$\|\nabla f_\lambda(\boldsymbol{x})\| = \lambda^{-1}\|\boldsymbol{x} - \operatorname{prox}_{\lambda f}(\boldsymbol{x})\| = \lambda^{-1}\|\boldsymbol{x} - \hat{\boldsymbol{x}}\|.$$

The second statement follows from the definition of Moreau envelope and proximal point:

$$f(\hat{\boldsymbol{x}}) + \frac{1}{2\lambda}\|\boldsymbol{x} - \hat{\boldsymbol{x}}\|^2 \le f(\boldsymbol{x}) + \frac{1}{2\lambda}\|\boldsymbol{x} - \boldsymbol{x}\|^2 = f(\boldsymbol{x}).$$

The last one is given by the optimality condition of proximal point and the item (i): $0 \in \partial f(\hat{\boldsymbol{x}}) + \frac{1}{\lambda}(\hat{\boldsymbol{x}} - \boldsymbol{x})$. ■

Remark 12. *A small gradient $\|\nabla f_\lambda(\boldsymbol{x})\|$ implies that x is "near" some point $\hat{\boldsymbol{x}}$ that is approximate stationary.*

3 Subgradient Methods for Weakly Convex Functions

Algorithm 1 Subgradient Methods for Weakly Convex Functions

```
For $k = 0$ to $K$ do
    $\boldsymbol{x}^{k+1} = \boldsymbol{x}^k - t_k \boldsymbol{g}_k, \quad$ where $\boldsymbol{g}_k \in \partial f(\boldsymbol{x}^k)$.
End
Sample $k^* \in \{0, 1, 2, \ldots, K\}$ according to the probability
    $\mathbb{P}(k^* = k) = \frac{t_k}{\sum_{j=0}^K t_j}$.
```

Theorem 13. *Suppose that f is ρ-weakly convex and L-Lipschitz. Let $\boldsymbol{x}_k^*$ be the point returned by the algorithm. Then, for any $\hat{\rho} > \rho$, we have*

$$\mathbb{E}\left[\|\nabla f_{\frac{1}{\hat{\rho}}}(\boldsymbol{x}^{k^*})\|^2\right] \leq \frac{\hat{\rho}}{\hat{\rho} - \rho} \frac{(f_{\frac{1}{\hat{\rho}}}(\boldsymbol{x}^0) - f^*) + \frac{\hat{\rho}L^2}{2} \sum_{k=0}^K t_k^2}{\sum_{k=0}^K t_k}$$

Remark 14. *The term $\frac{\hat{\rho}L^2}{2} \sum_{k=0}^K t_k^2$ is the price of non-smoothness. It is also worth mentioning that the same cost applies to the stochastic case. Therefore, both subgradient and stochastic subgradient methods will achieve the same convergence rate.*

Remark 15. *It is natural to set the step size as $t_k = \frac{1}{\sqrt{k}}$ as we hope the term $\sum_{k=0}^\infty t_k^2$ is summable but $\sum_{k=0}^\infty t_k^2$ is not.*

Proof. Set $\hat{\boldsymbol{x}}^k = \operatorname{prox}_{\frac{1}{\hat{\rho}} f}(\boldsymbol{x}^k)$. Then we have

$$\begin{aligned}
f_{\frac{1}{\hat{\rho}}}(\boldsymbol{x}^{k+1}) &= \inf_{\boldsymbol{y}} \left\{ f(\boldsymbol{y}) + \frac{\hat{\rho}}{2}\|\boldsymbol{y} - \boldsymbol{x}^{k+1}\|^2 \right\} \\
&\leq f(\hat{\boldsymbol{x}}^k) + \frac{\hat{\rho}}{2}\|\hat{\boldsymbol{x}}^k - \boldsymbol{x}^{k+1}\|^2 \\
&= f(\hat{\boldsymbol{x}}^k) + \frac{\hat{\rho}}{2}\|\boldsymbol{x}^k - t_k \boldsymbol{g}_k - \hat{\boldsymbol{x}}^k\|^2 \\
&\leq f(\hat{\boldsymbol{x}}^k) + \frac{\hat{\rho}}{2}\|\boldsymbol{x}^k - \hat{\boldsymbol{x}}^k\|^2 + \frac{\hat{\rho}}{2} t_k^2 L^2 + \hat{\rho} t_k \langle \boldsymbol{x}^k - \hat{\boldsymbol{x}}, \boldsymbol{g}_k \rangle \\
&= f_{\frac{1}{\hat{\rho}}}(\boldsymbol{x}^k) + \frac{\hat{\rho}}{2} t_k^2 L^2 + \hat{\rho} t_k \langle \boldsymbol{x}^k - \hat{\boldsymbol{x}}^k, \boldsymbol{g}_k \rangle \\
&\leq f_{\frac{1}{\hat{\rho}}}(\boldsymbol{x}^k) + \frac{\hat{\rho}}{2} t_k^2 L^2 + \hat{\rho}\left(f(\hat{\boldsymbol{x}}^k) - f(\boldsymbol{x}^k) + \frac{\rho}{2}\|\boldsymbol{x}^k - \hat{\boldsymbol{x}}^k\|^2 \right),
\end{aligned}$$

where the second inequality follows from the L-Lipschitz continuity of f, and the last one is given by the weakly convex of f.

Telescope this sum from $k = 0$ to $k = K$:

$$f_{\frac{1}{\hat{\rho}}}(\boldsymbol{x}^{K+1}) \leq f_{\frac{1}{\hat{\rho}}}(\boldsymbol{x}^0) + \frac{\hat{\rho}}{2} L^2 \sum_{k=0}^K t_k^2 - \hat{\rho} \sum_{k=0}^K t_k \left(f(\hat{\boldsymbol{x}}^k) - f(\boldsymbol{x}^k) + \frac{\rho}{2}\|\boldsymbol{x}^k - \hat{\boldsymbol{x}}^k\|^2 \right),$$

which is:

$$\frac{1}{\sum_{k=0}^K t_k} \sum_{k=0}^K t_k \left(f(\hat{\boldsymbol{x}}^k) - f(\boldsymbol{x}^k) + \frac{\rho}{2}\|\boldsymbol{x}^k - \hat{\boldsymbol{x}}^k\|^2 \right) \leq \frac{\left(f_{\frac{1}{\hat{\rho}}}(\boldsymbol{x}^0) - f_{\frac{1}{\hat{\rho}}}(\boldsymbol{x}^{k+1}) \right) + \frac{\hat{\rho}}{2} L^2 \sum_{k=0}^K t_k^2}{\hat{\rho} \sum_{k=0}^K t_k}.$$

By the strong convexity of $\boldsymbol{x} \to f(\boldsymbol{x}) + \frac{\hat{\rho}}{2}\|\boldsymbol{x} - \boldsymbol{x}^k\|^2$, we have that:

$$\begin{aligned} f(\boldsymbol{x}^k) - f(\hat{\boldsymbol{x}}^k) + \frac{\rho}{2}\|\boldsymbol{x}^k - \boldsymbol{x}^k\|^2 &= \left(f(\boldsymbol{x}^k) + \frac{\hat{\rho}}{2}\|\boldsymbol{x}^k - \hat{\boldsymbol{x}}^k\|^2\right) - \left(f(\hat{\boldsymbol{x}}^k) + \frac{\hat{\rho}}{2}\|\boldsymbol{x}^k - \hat{\boldsymbol{x}}^k\|^2\right) + \frac{\hat{\rho} - \rho}{2}\|\hat{\boldsymbol{x}}^k - \boldsymbol{x}^k\|^2 \\ &\geq (\hat{\rho} - \rho)\|\hat{\boldsymbol{x}}^k - \boldsymbol{x}^k\|^2 \\ &= \frac{\hat{\rho} - \rho}{\rho^2}\|\nabla f_{\frac{1}{\hat{\rho}}}(\boldsymbol{x}^k)\|^2. \end{aligned}$$

Thus, we have

$$\mathbb{E}\left[\|\nabla f_{\frac{1}{\hat{\rho}}}(\boldsymbol{x}_k^*)\|^2\right] \leq \frac{\hat{\rho}}{\hat{\rho} - \rho} \frac{(f_{\frac{1}{\hat{\rho}}}(\boldsymbol{x}^0) - f^*) + \frac{\hat{\rho}L^2}{2}\sum_{k=0}^K t_k^2}{\sum_{k=0}^K t_k}.$$

■

4 Proximal Gradient Descent

Then, we continue to consider the structured nonsmooth probelm as below:

$$\inf_{\boldsymbol{x} \in \mathbb{R}^d} F(\boldsymbol{x}) := f(\boldsymbol{x}) + g(\boldsymbol{x}) \tag{CP}$$

where:

(i) $f : \mathbb{R}^d \to \mathbb{R}$ is L-smooth,

(ii) $g : \mathbb{R}^d \to \overline{\mathbb{R}}$ is convex, closed, and (possibly) non-smooth,

(iii) prox_{tg} is easily computed. (e.g., ℓ_1, ℓ_∞ norms)

Example 16 (LASSO Problem)**.**

$$\min_{\boldsymbol{x} \in \mathbb{R}^d} \left[\frac{1}{2}\|A\boldsymbol{x} - b\|^2 + \lambda\|\boldsymbol{x}\|_1\right]$$

The proximal operator for the ℓ_1 norm is given by:

$$\operatorname{prox}_{\lambda\|\boldsymbol{x}\|_1}(\boldsymbol{x}) = \arg\min_{\boldsymbol{y}} \left[\frac{1}{2}\|\boldsymbol{y} - \boldsymbol{x}\|^2 + \lambda\|\boldsymbol{y}\|_1\right]$$

This is also known as soft-thresholding *and satisfies the Karush-Kuhn-Tucker (KKT) conditions.*

Remark 17. *(i) If $g = 0$, then the PGD reduces to the GD.*

(ii) If $f = 0$, then the PGD reduces to the proximal point descent.

(iii) If $g = \mathbf{1}_{\mathcal{X}}$ (indicator of set $\mathcal{X}$), then PGD becomes projected GD.

References

[1] Damek Davis and Dmitriy Drusvyatskiy. Stochastic subgradient method converges at the rate $\mathcal{O}(k^{-1/4})$ on weakly convex functions. *arXiv preprint arXiv:1802.02988*, 2018. 1

[2] Damek Davis and Dmitriy Drusvyatskiy. Stochastic model-based minimization of weakly convex functions. *SIAM Journal on Optimization*, 29(1):207–239, 2019. 1

[3] R. Tyrrell Rockafellar and Roger J-B Wets. *Variational Analysis*, volume 317. Springer Science & Business Media, 2009. 2

Algorithm 2

The Gradient Descent (GD) follows:

$$\begin{aligned} \boldsymbol{x}^{k+1} &= \boldsymbol{x}^k - t\nabla f(\boldsymbol{x}^k) \\ &= \arg\min_{\boldsymbol{x}} \left\{ \nabla f(\boldsymbol{x}^k)^\top (\boldsymbol{x} - \boldsymbol{x}^k) + \frac{1}{2t}\|\boldsymbol{x} - \boldsymbol{x}^k\|^2 \right\}, \end{aligned} \tag{GD}$$

where the minimization objective is the quadratic upper bound.
The Promixal Gradient Descent (GD) follows:

$$\begin{aligned} \boldsymbol{x}^{k+1} &= \arg\min_{\boldsymbol{x}} \left\{ \nabla f(\boldsymbol{x}^k)^\top (\boldsymbol{x} - \boldsymbol{x}^k) + \frac{1}{2t}\|\boldsymbol{x} - \boldsymbol{x}^k\|^2 + g(\boldsymbol{x}) \right\} \\ &= \arg\min_{\boldsymbol{x}} \left\{ \frac{1}{2t}\|\boldsymbol{x} - (\boldsymbol{x}^k - t\nabla f(\boldsymbol{x}^k))\|^2 + g(\boldsymbol{x}) \right\} \\ &= \mathrm{prox}_{tg}(\boldsymbol{x}^k - t\nabla f(\boldsymbol{x}^k)). \end{aligned} \tag{PGD}$$