CMPT 409/981: Optimization for Machine Learning

Lecture 13

Sharan Vaswani

October 31, 2022

Recap

Function class	L-smooth + convex	L-smooth + μ-strongly convex
GD	$O\left({n}/{\epsilon}\right)$	$O\left(n\,\kappa \log\left({1}/{\epsilon}\right)\right)$
Nesterov Acceleration	$O\left({n}/{\sqrt{\epsilon}}\right)$	$O\left(n\sqrt{\kappa} \log\left({1}/{\epsilon}\right)\right)$
SGD	$O\left({1}/{\epsilon^2}\right)$	$O\left({1}/{\epsilon}\right)$
SGD under exact interpolation	$O\left({1}/{\epsilon}\right)$	$O\left(\kappa \log\left({1}/{\epsilon}\right)\right)$
Variance reduced methods (SVRG [JZ13], SARAH [NLST17])	$O\left(\left(n + {1}/{\epsilon}\right)\log({1}/{\epsilon})\right)$	$O\left(\left(n + \kappa\right)\log\left({1}/{\epsilon}\right)\right)$
Accelerated variance reduced methods (Katyusha [AZ17], Varag [LLZ19]),	$O\left(\left(n + {1}/{\sqrt{\epsilon}}\right)\log({1}/{\epsilon})\right)$	$O\left(\left(n + \sqrt{\kappa}\right)\log\left({1}/{\epsilon}\right)\right)$

Table 1: Number of gradient evaluations for obtaining an ϵ-sub-optimality when minimizing a finite-sum.

Today, we will look at minimizing non-smooth, but Lipschitz (strongly)-convex functions.

Lipschitz Functions

Recall that for Lipschitz functions, for all $x, y \in \mathcal{D}$, there exists a constant $G < \infty$,

$$|f(y) - f(x)| \leq G \|x - y\| \ .$$

This immediately implies that the gradients are bounded, i.e. for all $w \in \mathcal{D}$, $\|\nabla f(w)\| \leq G$.

Example: Hinge loss: $f(w) = \max\{0, 1 - y\langle w, x\rangle\}$ is Lipschitz with $G = \|y\, x\|$

Compare this to smooth functions that satisfy $\|\nabla f(x) - \nabla f(y)\| \leq L \|x - y\|$. Lipschitz functions are not necessarily smooth, and smooth functions are not necessarily Lipschitz.

Example: $f(w) = |w|$ is 1-Lipschitz, but not smooth (gradient changes from -1 to $+1$ at $w = 0$). On the other hand, $f(w) = \frac{1}{2}\|x\|_2^2$ is 1-smooth, but not Lipschitz (the gradient is equal to x and hence not bounded).

Subgradients

Subgradient: For a convex function f, the subgradient of f at $x \in \mathcal{D}$ is a vector g that satisfies the inequality,

$$f(y) \geq f(x) + \langle g, y - x \rangle$$

This is similar to the first-order definition of convexity, with the subgradient instead of the gradient. Importantly, the subgradient is not unique.

Example: For $f(w) = |w|$ at $w = 0$, vectors with slope $[-1, 1]$ and passing through the origin are subgradients.

Subdifferential: Set of subgradients of f at $w \in \mathcal{D}$ is referred to as the subdifferential and denoted by $\partial f(w)$. Formally, $\partial f(w) = \{g | \forall y \in \mathcal{D}; f(y) \geq f(x) + \langle g, y - x \rangle\}$.

For $f : \mathcal{D} \to \mathbb{R}$, iff $\forall w \in \mathcal{D}$, $\partial f(w) \neq \emptyset$, f is convex. If f is convex and differentiable at w, then $\nabla f(w) \in \partial f(w)$ (see [B$^+$15, Proposition 1.1] for a proof)

Subgradients

Example: For $f(w) = |w|$,

$$\partial f(w) = \begin{cases} \{-1\} & \text{for } x > 0 \\ [-1, 1] & \text{for } x = 0 \\ \{1\} & \text{for } x < 0 \end{cases}$$

Q: Compute the subdifferential for the Hinge loss $f(w) = \max\{0, 1 - \langle z, w \rangle\}$

Subgradient Descent

Algorithmically, we can use the subgradient instead of the gradient in gradient descent, and use the resulting algorithm to minimize Lipschitz functions.

Projected Subgradient Descent: $w_{k+1} = \Pi_{\mathcal{D}}\left[w_k - \eta_k g_k\right]$, where $g_k \in \partial f(w_k)$.

Similar to GD, we can interpret subgradient descent as:

$$w_{k+1} = \operatorname*{arg\,min}_{w \in \mathcal{D}} \left[\langle g_k, w\rangle + \frac{1}{2\eta_k}\left\|w - w_k\right\|^2\right]$$

Unlike for smooth, convex functions, we cannot relate the subgradient norm to the suboptimality in the function values. **Example**: For $f(w) = |w|$, for all $w > 0$ (including $w = 0^+$), $\|g\| = 1$.

Consequently, in order to converge to the minimizer, we need to explicitly decrease the step-size resulting in slower convergence. E.g., for Lipschitz, convex functions, $\eta_k = O(1/\sqrt{k})$ and subgradient descent will result in $\Theta\left(\frac{1}{\sqrt{T}}\right)$ convergence.

Minimizing convex, Lipschitz functions using Subgradient Descent

For simplicity, let us assume that $\mathcal{D} = \mathbb{R}^d$ and analyze the convergence of subgradient descent.

Claim: For G-Lipschitz, convex functions, for $\eta > 0$, T iterations of subgradient descent with $\eta_k = \eta/\sqrt{k}$ satisfies the following convergence, where $\bar{w}_T = \sum_{k=0}^{T-1} w_k/T$,

$$f(\bar{w}_T) - f(w^*) \leq \frac{1}{\sqrt{T}} \left[\frac{\|w_0 - w^*\|^2}{2\eta} + \frac{G^2 \eta \, [1 + \log(T)]}{2} \right].$$

Proof: Similar to the previous proofs, using the update $w_{k+1} = w_k - \eta_k g_k$ where $g_k \in \partial f(w_k)$,

$$\begin{aligned}
\|w_{k+1} - w^*\|^2 &= \|w_k - w^*\|^2 - 2\eta_k \langle g_k, w_k - w^* \rangle + \eta_k^2 \|g_k\|^2 \\
&\leq \|w_k - w^*\|^2 - 2\eta_k [f(w_k) - f(w^*)] + \eta_k^2 \|g_k\|^2 \\
&\hspace{8em} \text{(Definition of subgradient with } x = w_k,\ y = w^*\text{)} \\
&\leq \|w_k - w^*\|^2 - 2\eta_k [f(w_k) - f(w^*)] + \eta_k^2 G^2 \\
&\hspace{8em} \text{(Since } f \text{ is } G\text{-Lipschitz)}
\end{aligned}$$

$$\implies \eta_k [f(w_k) - f(w^*)] \leq \frac{\|w_k - w^*\|^2 - \|w_{k+1} - w^*\|^2}{2} + \frac{\eta_k^2 G^2}{2}$$

Minimizing convex, Lipschitz functions using Subgradient Descent

Recall that $\eta_k[f(w_k) - f(w^*)] \leq \frac{\|w_k - w^*\|^2 - \|w_{k+1} - w^*\|^2}{2} + \frac{\eta_k^2 G^2}{2}$,

$$\implies \eta_{\min} \sum_{k=0}^{T-1} [f(w_k) - f(w^*)] \leq \sum_{k=0}^{T-1} \left[\frac{\|w_k - w^*\|^2 - \|w_{k+1} - w^*\|^2}{2} \right] + \frac{G^2}{2} \sum_{k=0}^{T-1} \eta_k^2$$

$$\leq \frac{\|w_0 - w^*\|^2}{2} + \frac{G^2}{2} \sum_{k=0}^{T-1} \eta_k^2$$

$$\implies \frac{\eta}{\sqrt{T}} \sum_{k=0}^{T-1} [f(w_k) - f(w^*)] \leq \frac{\|w_0 - w^*\|^2}{2} + \frac{G^2 \eta^2}{2} \sum_{k=0}^{T-1} \frac{1}{k} \qquad \text{(Since } \eta_k = \eta/\sqrt{k}\text{)}$$

$$\implies \frac{\sum_{k=0}^{T-1} [f(w_k) - f(w^*)]}{T} \leq \frac{1}{\sqrt{T}} \left[\frac{\|w_0 - w^*\|^2}{2\eta} + \frac{G^2 \eta \, [1 + \log(T)]}{2} \right]$$

$$\implies f(\bar{w}_T) - f(w^*) \leq \frac{1}{\sqrt{T}} \left[\frac{\|w_0 - w^*\|^2}{2\eta} + \frac{G^2 \eta \, [1 + \log(T)]}{2} \right]$$

(Using Jensen's inequality on the LHS, and by definition of $\bar{w}_T$.)

Minimizing convex, Lipschitz functions using Subgradient Descent

Recall that $f(\bar{w}_T) - f(w^*) \leq \frac{1}{\sqrt{T}} \left[\frac{\|w_0 - w^*\|^2}{2\eta} + \frac{G^2 \eta \, [1 + \log(T)]}{2}\right]$. The above proof works for any value of η and we can modify the proof to set the "best" value of η.

For this, let us use a constant step-size $\eta_k = \eta$. Following the same proof as before,

$$\eta_{\min} \sum_{k=0}^{T-1} [f(w_k) - f(w^*)] \leq \frac{\|w_0 - w^*\|^2}{2} + \frac{G^2}{2} \sum_{k=0}^{T-1} \eta_k^2$$

$$\implies \sum_{k=0}^{T-1} [f(w_k) - f(w^*)] \leq \frac{\|w_0 - w^*\|^2}{2\eta} + \frac{G^2 T \eta}{2} \qquad \text{(Since } \eta_k = \eta\text{)}$$

Setting $\eta = \frac{\|w_0 - w^*\|}{G\sqrt{T}}$, dividing by T and using Jensen's inequality on the LHS,

$$f(\bar{w}_T) - f(w^*) \leq \frac{G \|w_0 - w^*\|}{\sqrt{T}}$$

For Lipschitz, convex functions, the above $O(1/\epsilon^2)$ rate is optimal, but we require knowledge of $G, \|w_0 - w^*\|, T$ to set the step-size.

Minimizing convex, Lipschitz functions using Subgradient Descent

Recall that for smooth, convex functions, we could use Nesterov acceleration to obtain a faster $O(1/\sqrt{\epsilon})$ rate. On the other hand, for Lipschitz, convex functions, subgradient descent is optimal.

In order to get the $\frac{G\|w_0 - w^*\|}{\sqrt{T}}$ rate, we needed knowledge of G and $\|w_0 - w^*\|$ to set the step-size. There are various techniques to set the step-size in an adaptive manner.

- AdaGrad [DHS11] is adaptive to G, but still requires knowing a quantity related $\|w_0 - w^*\|$ to select the "best" step-size. This influences the practical performance of AdaGrad.
- Polyak step-size [HK19] attains the desired rate without knowledge of G or $\|w_0 - w^*\|$, but requires knowing f^*.
- Coin-Betting [OP16] does not require knowledge of $\|w_0 - w^*\|$. It only requires an estimate of G and is robust to its misspecification in theory (but not quite in practice).

Minimizing convex, Lipschitz functions using Subgradient Descent

For Lipschitz, strongly-convex functions, subgradient descent attains an $\Theta\left(\frac{1}{\epsilon}\right)$ rate. For this, the step-size depends on μ and the proof is similar to the one on (Slide 6, Lecture 10).

Subgradient descent is also optimal for Lipschitz, strongly-convex functions.

For Lipschitz functions, the convergence rates for SGD are the same as GD (with similar proofs).

Function class	L-smooth + convex	L-smooth + μ-strongly convex	G-Lipschitz + convex	G-Lipschitz + μ-strongly convex
GD	$O\left(1/\epsilon\right)$	$O\left(\kappa \log\left(1/\epsilon\right)\right)$	$\Theta\left(1/\epsilon^2\right)$	$\Theta\left(1/\epsilon\right)$
SGD	$\Theta\left(1/\epsilon^2\right)$	$\Theta\left(1/\epsilon\right)$	$\Theta\left(1/\epsilon^2\right)$	$\Theta\left(1/\epsilon\right)$

Table 2: Number of iterations required for obtaining an ϵ-sub-optimality.

References i

- Zeyuan Allen-Zhu, *Katyusha: The first direct acceleration of stochastic gradient methods*, The Journal of Machine Learning Research **18** (2017), no. 1, 8194–8244.
- Sébastien Bubeck et al., *Convex optimization: Algorithms and complexity*, Foundations and Trends® in Machine Learning **8** (2015), no. 3-4, 231–357.
- John Duchi, Elad Hazan, and Yoram Singer, *Adaptive subgradient methods for online learning and stochastic optimization.*, Journal of machine learning research **12** (2011), no. 7.
- Elad Hazan and Sham Kakade, *Revisiting the polyak step size*, arXiv preprint arXiv:1905.00313 (2019).
- Rie Johnson and Tong Zhang, *Accelerating stochastic gradient descent using predictive variance reduction*, Advances in neural information processing systems **26** (2013).

- Guanghui Lan, Zhize Li, and Yi Zhou, *A unified variance-reduced accelerated gradient method for convex optimization*, Advances in Neural Information Processing Systems **32** (2019).
- Lam M Nguyen, Jie Liu, Katya Scheinberg, and Martin Takáč, *Sarah: A novel method for machine learning problems using stochastic recursive gradient*, International Conference on Machine Learning, PMLR, 2017, pp. 2613–2621.
- Francesco Orabona and Dávid Pál, *Coin betting and parameter-free online learning*, Advances in Neural Information Processing Systems **29** (2016).