CMPT 409/981: Optimization for Machine Learning

Lecture 13

Sharan Vaswani

October 24, 2024

Minimizing smooth, strongly-convex functions

For minimizing smooth, strongly-convex functions $f(w) = \frac{1}{n}\sum_{i=1}^{n} f_i(w)$ to an ϵ-suboptimality,

- Deterministic GD requires $O(\kappa \log(1/\epsilon))$ iterations, and $O(n\, \kappa \log(1/\epsilon))$ gradient evaluations.
- SGD with a decreasing step-size requires $O(1/\epsilon)$ iterations, and $O(1/\epsilon)$ gradient evaluations.
- Under exact interpolation, SGD with a constant step-size requires $O(\kappa \log(1/\epsilon))$ iterations, and $O(\kappa \log(1/\epsilon))$ gradient evaluations.
- For finite-sum problems of the form $\frac{1}{n}\sum_{i=1}^{n} f_i(w)$, **variance reduced methods** require $O((n + \kappa) \log(1/\epsilon))$ gradient evaluations.

Variance Reduced Methods

- Recall that under exact interpolation, the variance decreases as we approach the minimizer.
- In contrast, variance reduced (VR) methods explicitly reduce the variance by either storing the past stochastic gradients to approximate the full gradient [SLRB17] or by computing the full gradient every "few" iterations [JZ13].
- VR methods only require f to be a finite sum, and make no interpolation assumption.
- With variance reduction, we can use acceleration techniques to improve the dependence on the condition number, and require $O((n + \sqrt{\kappa}) \log(1/\epsilon))$ gradient evaluations [AZ17].
- For smooth, convex finite-sum problems, variance reduced techniques require $O\left((n + \frac{1}{\epsilon}) \log(1/\epsilon)\right)$ gradient evaluations [NLST17], compared to deterministic GD that requires $O(\frac{n}{\epsilon})$ gradient evaluations and SGD that requires $O(\frac{1}{\epsilon^2})$ gradient evaluations.
- We will use SVRG (Stochastic Variance Reduced Gradient) [JZ13] for smooth, strongly-convex finite-sum problems, and prove that it requires $O((n + \kappa) \log(1/\epsilon))$ gradient evaluations.

SVRG

For simplicity, we will use Loopless SVRG [KHR20] that has a simpler implementation and analysis compared to the original paper [JZ13].

Algorithm SVRG

function SVRG $(f,\ w_0,\ \eta,\ p \in (0,1])$

$v_0 = w_0$

for $k = 0, \ldots, T-1$ **do**

  $g_k = \nabla f_{ik}(w_k) - \nabla f_{ik}(v_k) + \nabla f(v_k)$

  $w_{k+1} = w_k - \eta g_k$

  $v_{k+1} = \begin{cases} v_k \text{ with probability } 1-p \\ w_k \text{ with probability } p \end{cases}$

end for

return w_T

Minimizing smooth, strongly-convex functions using SVRG

Claim: When minimizing $f(w) = \frac{1}{n}\sum_{i=1}^{n} f_i(w)$ such that (i) f is μ-strongly convex, (ii) each f_i is convex and L-smooth, T iterations of SVRG with $\eta = \frac{1}{6L}$ and $p = \frac{1}{n}$ returns iterate w_T,

$$\mathbb{E}[\|w_T - w^*\|^2] \leq \left(\max\left\{\left(1 - \frac{\mu}{6L}\right), \left(1 - \frac{1}{2n}\right)\right\}\right)^T \left[2n\,\|w_0 - w^*\|^2\right].$$

Case 1: $(1 - \frac{\mu}{6L}) \leq (1 - \frac{1}{2n}) \implies n \geq 3\kappa$. In this case, for achieving an ϵ-suboptimality, we need T iterations such that $T \geq 2n \log\left(\frac{2n\,\|w_0 - w^*\|^2}{\epsilon}\right)$.

Case 2: $(1 - \frac{\mu}{6L}) > (1 - \frac{1}{2n}) \implies n \leq 3\kappa$. In this case, for achieving an ϵ-suboptimality, we need T iterations such that $T \geq 6\kappa \log\left(\frac{2n\,\|w_0 - w^*\|^2}{\epsilon}\right)$.

- Putting cases together, for achieving an ϵ-suboptimality, we need $T = O\left((n + \kappa)\log(1/\epsilon)\right)$.

- In each iteration, the number of expected gradient evaluations is $(1 - p)(2) + (p)(n + 2) = pn + 2 = 3$. Hence, in expectation, SVRG requires $O\left((n + \kappa)\log(1/\epsilon)\right)$ gradient evaluations to achieve an ϵ-suboptimality.

Minimizing smooth, strongly-convex functions using SVRG

Proof: Using the algorithm update, $w_{k+1} = w_k - \eta g_k$ and following a similar proof as before,

$$\|w_{k+1} - w^*\|^2 = \|w_k - w^*\|^2 - 2\eta \langle g_k, w_k - w^* \rangle + \eta^2 \|g_k\|^2$$

$$\implies \mathbb{E} \|w_{k+1} - w^*\|^2 = \|w_k - w^*\|^2 - 2\eta \langle \mathbb{E}[g_k], w_k - w^* \rangle + \eta^2 \, \mathbb{E}[\|g_k\|^2]$$

(Since η does not depend on i_k)

$$= \|w_k - w^*\|^2 - 2\eta \langle \nabla f(w_k), w_k - w^* \rangle + \eta^2 \, \mathbb{E}[\|g_k\|^2]$$

$$(\mathbb{E}[g_k] = \mathbb{E}[\nabla f_{ik}(w_k) - \nabla f_{ik}(v_k) + \nabla f(v_k)] = \nabla f(w_k))$$

By strong-convexity,

$$\mathbb{E} \|w_{k+1} - w^*\|^2 \leq (1 - \mu\eta) \|w_k - w^*\|^2 - 2\eta \, [f(w_k) - f(w^*)] + \eta^2 \, \mathbb{E}[\|g_k\|^2] \qquad (1)$$

Next, we will bound $\mathbb{E}[\|g_k\|^2]$.

Minimizing smooth, strongly-convex functions using SVRG

$$\mathbb{E}[\|g_k\|^2] = \mathbb{E}[\|\nabla f_{ik}(w_k) - \nabla f_{ik}(v_k) + \nabla f(v_k)\|^2]$$
$$= \mathbb{E}[\|\nabla f_{ik}(w_k) - \nabla f_{ik}(w^*) + \nabla f_{ik}(w^*) - \nabla f_{ik}(v_k) + \nabla f(v_k)\|^2]$$
$$\leq 2\mathbb{E}\left[\|\nabla f_{ik}(w_k) - \nabla f_{ik}(w^*)\|^2\right] + 2\mathbb{E}\left[\|\nabla f_{ik}(w^*) - \nabla f_{ik}(v_k) + \nabla f(v_k)\|^2\right]$$
$$((a+b)^2 \leq 2a^2 + 2b^2)$$
$$= 2\mathbb{E}\left[\|\nabla f_{ik}(w_k) - \nabla f_{ik}(w^*)\|^2\right] + 2\mathbb{E}\left[\|\nabla f_{ik}(w^*) - \nabla f_{ik}(v_k) - \mathbb{E}\left[\nabla f_{ik}(w^*) - \nabla f_{ik}(v_k)\right]\|^2\right]$$
$$(\text{Since } \mathbb{E}[\nabla f_{ik}(w^*)] = \nabla f(w^*) = 0)$$

For any vector x, $\mathbb{E}\left[\|x - \mathbb{E}[x]\|^2\right] \leq \mathbb{E}[\|x\|^2]$. Using this with $x = \nabla f_{ik}(w^*) - \nabla f_{ik}(v_k)$

$$\leq 2\mathbb{E}\left[\|\nabla f_{ik}(w_k) - \nabla f_{ik}(w^*)\|^2\right] + 2\mathbb{E}\left[\|\nabla f_{ik}(w^*) - \nabla f_{ik}(v_k)\|^2\right]$$
$$\leq 4L\,\mathbb{E}\left[f_{ik}(w_k) - f_{ik}(w^*) + \langle \nabla f_{ik}(w^*), w^* - w_k\rangle\right] + 2\mathbb{E}\left[\|\nabla f_{ik}(w^*) - \nabla f_{ik}(v_k)\|^2\right]$$
$$(\text{Smoothness of } f_{ik})$$

$$\implies \mathbb{E}[\|g_k\|^2] \leq 4L\,\mathbb{E}[f(w_k) - f(w^*)] + 2\mathbb{E}\left[\|\nabla f_{ik}(w^*) - \nabla f_{ik}(v_k)\|^2\right] \quad (2)$$

Minimizing smooth, strongly-convex functions using SVRG

Using eq. (1) with eq. (2),

$$\begin{aligned}\mathbb{E}\left\|w_{k+1}-w^*\right\|^2 &\leq (1-\mu\eta)\left\|w_k-w^*\right\|^2-2\eta\left[f(w_k)-f(w^*)\right] \\ &+\eta^2\left[4L\,\mathbb{E}[f(w_k)-f(w^*)]+2\mathbb{E}\left[\left\|\nabla f_{ik}(w^*)-\nabla f_{ik}(v_k)\right\|^2\right]\right] \\ &\leq (1-\mu\eta)\left\|w_k-w^*\right\|^2+(4L\eta^2-2\eta)\,\mathbb{E}\left[f(w_k)-f(w^*)\right] \\ &+\frac{2\eta^2}{n}\sum_{i=1}^{n}\left[\left\|\nabla f_i(w^*)-\nabla f_i(v_k)\right\|^2\right]\end{aligned}$$

Define $\mathcal{D}_k := \frac{4\eta^2}{pn}\sum_{i=1}^n\left[\left\|\nabla f_i(w^*)-\nabla f_i(v_k)\right\|^2\right]$.

$$\mathbb{E}\left\|w_{k+1}-w^*\right\|^2 \leq (1-\mu\eta)\left\|w_k-w^*\right\|^2+(4L\eta^2-2\eta)\,\mathbb{E}\left[f(w_k)-f(w^*)\right]+\frac{p}{2}\mathcal{D}_k \qquad (3)$$

Minimizing smooth, strongly-convex functions using SVRG

Recall that $\mathcal{D}_k = \frac{4\eta^2}{pn} \sum_{i=1}^{n} \left[\left\|\nabla f_i(w^*) - \nabla f_i(v_k)\right\|^2\right]$. Using the algorithm,

$$\mathbb{E}[\mathcal{D}_{k+1}] = (1-p)\mathcal{D}_k + p\,\frac{4\eta^2}{pn} \sum_{i=1}^{n} \left[\left\|\nabla f_i(w^*) - \nabla f_i(w_k)\right\|^2\right]$$

$$\leq (1-p)\mathcal{D}_k + \frac{8\eta^2\,L}{n} \sum_{i=1}^{n} \left[f_i(w_k) - f_i(w^*) + \langle \nabla f_i(w^*), w^* - w_k\rangle\right] \qquad \text{(Smoothness)}$$

$$\implies \mathbb{E}[\mathcal{D}_{k+1}] \leq (1-p)\mathcal{D}_k + 8\eta^2\,L\,\left[f(w_k) - f(w^*)\right] \qquad (4)$$

Minimizing smooth, strongly-convex functions using SVRG

Using eq. (3) + eq. (4),

$$
\begin{aligned}
\mathbb{E}\left\|w_{k+1}-w^*\right\|^2+\mathbb{E}[\mathcal{D}_{k+1}] &\leq (1-\mu\eta)\left\|w_k-w^*\right\|^2+(4L\eta^2-2\eta)\,\mathbb{E}\left[f(w_k)-f(w^*)\right]+\frac{p}{2}\mathcal{D}_k \\
&\quad +(1-p)\mathcal{D}_k+8\eta^2 L\,[f(w_k)-f(w^*)] \\
&= (1-\mu\eta)\left\|w_k-w^*\right\|^2+(12L\eta^2-2\eta)\,[f(w_k)-f(w^*)]+\left(1-\frac{p}{2}\right)\mathcal{D}_k \\
&= \left(1-\frac{\mu}{6L}\right)\left\|w_k-w^*\right\|^2+\left(1-\frac{p}{2}\right)\mathcal{D}_k \qquad \left(\text{Since } \eta=\tfrac{1}{6L}\right) \\
&\leq \max\left\{\left(1-\frac{\mu}{6L}\right),\left(1-\frac{p}{2}\right)\right\}\left[\left\|w_k-w^*\right\|^2+\mathcal{D}_k\right]
\end{aligned}
$$

$$
\mathbb{E}\left[\left\|w_{k+1}-w^*\right\|^2+\mathcal{D}_{k+1}\right] \leq \max\left\{\left(1-\frac{\mu}{6L}\right),\left(1-\frac{1}{2n}\right)\right\}\left[\left\|w_k-w^*\right\|^2+\mathcal{D}_k\right] \qquad \left(\text{Since } p=\tfrac{1}{n}\right)
$$

Define $\Phi_k := \left[\left\|w_k-w^*\right\|^2+\mathcal{D}_k\right]$ and $\rho := \max\left\{\left(1-\frac{\mu}{6L}\right),\left(1-\frac{1}{2n}\right)\right\}$

$$
\implies \mathbb{E}[\Phi_{k+1}] \leq \rho\,\Phi_k
$$

Minimizing smooth, strongly-convex functions using SVRG

Recall that $\mathbb{E}[\Phi_{k+1}] \leq \rho \Phi_k$. Taking expectation w.r.t the randomness in iterations from $k = 0$ to $T - 1$ and recursing,

$$
\begin{aligned}
\mathbb{E}[\Phi_T] &\leq \rho^T \Phi_0 \\
\implies \mathbb{E}[\|w_T - w^*\|^2] &\leq \rho^T \left[\|w_0 - w^*\|^2 + \mathcal{D}_0\right] && \text{(Lower bounding } \phi_T \text{ since } \mathcal{D}_T \text{ is positive)} \\
&= \rho^T \left[\|w_0 - w^*\|^2 + 4\eta^2 \sum_{i=1}^{n} \|\nabla f_i(w_0) - \nabla f_i(w^*)\|^2\right] \\
&\leq \rho^T \left[\|w_0 - w^*\|^2 + 4\eta^2 L^2 \sum_{i=1}^{n} \|w_0 - w^*\|^2\right] && \text{(Smoothness)} \\
\implies \mathbb{E}[\|w_T - w^*\|^2] &\leq \left(\max\left\{\left(1 - \frac{\mu}{6L}\right), \left(1 - \frac{1}{2n}\right)\right\}\right)^T \left[2n \|w_0 - w^*\|^2\right]
\end{aligned}
$$

(Since $\eta = \frac{1}{6L}$)

Questions?

Summary

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.

The final class of functions we will look at is 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, $f(w) = \frac{1}{2}\|w\|_2^2$ 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 for all y,

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

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 in $[-1, 1]$ and passing through the origin are subgradients.

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

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 $f(w) = |w|$,

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

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

Ans:

$$\partial f(w) = \begin{cases} \{0\} & \text{for } 1 - \langle z, w \rangle < 0 \\ \{-\alpha z | \alpha \in [0, 1]\} & \text{for } 1 - \langle z, w \rangle = 0 \\ \{-z\} & \text{for } 1 - \langle z, w \rangle > 0 \end{cases}$$

Subgradients

- For unconstrained minimization of convex, non-smooth functions, w^* is the minimizer of f iff $0 \in \partial f(w^*)$ (this is analogous to the smooth case).

Using the subgradient definition at $x = w^*$, if $0 \in \partial f(w^*)$, then, for all y,

$$f(y) \geq f(w^*) + \langle 0, y - w^* \rangle \implies f(y) \geq f(w^*),$$

and hence w^* is a minimizer of f.

Example: For $f(w) = |w|$, $0 \in \partial f(0)$ and hence $w^* = 0$.

Similarly, when minimizing convex, non-smooth functions over a constrained domain, if $w^* = \arg\min_{\mathcal{D}} f(w)$ iff $\exists g \in \partial f(w^*)$ such that $y \in \mathcal{D}$, $\langle g, y - w^* \rangle \geq 0$.

Subgradient Descent

- Algorithmically, we can use the subgradient instead of the gradient in GD, and use the resulting algorithm to minimize convex, 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:

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

- 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 $w = 0^+$), $\|g\| = 1$.

- Since the sub-gradient norm does not necessarily decrease closer to the solution, to converge to the minimizer, we need to explicitly decrease the step-size resulting in slower convergence.

Example: For Lipschitz, convex functions, $\eta_k = O(1/\sqrt{k})$ and subgradient descent will result in $\Theta\left(1/\sqrt{T}\right)$ convergence.

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.
- Rie Johnson and Tong Zhang, *Accelerating stochastic gradient descent using predictive variance reduction*, Advances in neural information processing systems **26** (2013).
- Dmitry Kovalev, Samuel Horváth, and Peter Richtárik, *Don't jump through hoops and remove those loops: Svrg and katyusha are better without the outer loop*, Algorithmic Learning Theory, PMLR, 2020, pp. 451–467.
- 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.
- Mark Schmidt, Nicolas Le Roux, and Francis Bach, *Minimizing finite sums with the stochastic average gradient*, Mathematical Programming **162** (2017), no. 1, 83–112.