Optimization for Data Science

Lec 12: Stochastic Gradient

Yaoliang Yu

UNIVERSITY OF WATERLOO | FACULTY OF MATHEMATICS
DAVID R. CHERITON SCHOOL OF COMPUTER SCIENCE

Problem

Minimization problem:

$$\min_{\mathbf{w} \in C} f(\mathbf{w})$$

- f smooth or subdifferentiable
- $C \subseteq \mathbb{R}^d$ a convex set
- Can only afford a noisy gradient

Where Is the Noise From?

- Measurement error
- Numerical error
- Scale constraint: most ML problems minimize an averaged loss

$$f(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n} \ell_i(\mathbf{w}), \quad \partial f(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n} \partial \ell_i(\mathbf{w})$$

- Convenience: objective can be reformulated as an expectation

$$f(\mathbf{w}) := \mathbb{E}_{\boldsymbol{\xi}}[f(\mathbf{w}, \boldsymbol{\xi})], \quad \partial f(\mathbf{w}) = \mathbb{E}_{\boldsymbol{\xi}}[\partial_{\mathbf{w}} f(\mathbf{w}, \boldsymbol{\xi})]$$

- Regularization: adding noise during training is common in ML
- Privacy: corrupt gradient with noise so that no one can infer user data

Algorithm 1: Stochastic Gradient

Input: $\mathbf{w}_0 \in \operatorname{dom} f$

for $t = 0, 1, 2, \ldots$ **do**

  choose step size η_t

  compute stochastic gradient $\mathbf{g}_t \leftarrow \nabla f(\mathbf{w}_t, \boldsymbol{\xi}_t)$

  $\mathbf{w}_{t+1} \leftarrow \mathrm{P}_C(\mathbf{w}_t - \eta_t \mathbf{g}_t)$

  $\mathbf{z}_t \leftarrow \sum_{k=0}^{t} \bar{\eta}_{t,k} \mathbf{w}_k$ `// ergodic averaging,` $\bar{\eta}_{t,k} := \eta_k / H_t, \;\; H_t := \sum_{k=0}^{t} \eta_k$

- For simplicity, assume stochastic gradient is unbiased, i.e.

$$\mathbb{E}_{\boldsymbol{\xi}_t}[\nabla f(\mathbf{w}_t, \boldsymbol{\xi}_t)] = \nabla f(\mathbf{w}_t)$$

- In general, step size $\eta_t \to 0$ (or $\sum_t \eta_t^2 < \infty$) and $\sum_t \eta_t = \infty$
- Surprisingly similar to the subgradient algorithm (including the analysis)

H. Robbins and S. Monro. "A Stochastic Approximation Method". *Annals of Mathematical Statistics*, vol. 22, no. 3 (1951), pp. 400–407.

Necessity of Diminishing Step Size

- Suppose we are at the minimizer $\mathbf{w}_\star$
- Gradient vanishes at $\nabla f(\mathbf{w}_\star)$
- But stochastic gradient $\nabla f(\mathbf{w}_\star, \xi)$ need not be zero
- With a non-vanishing step size, we will wander around $\mathbf{w}_\star$

Example

Algorithm 2: Perceptron

Input: Dataset $\mathcal{D} = \{\!\!\{(\mathbf{x}_i, \mathsf{y}_i) \in \mathbb{R}^d \times \{\pm 1\} : i = 1, \ldots, n\}\!\!\}$, initialization $\mathbf{w} \in \mathbb{R}^d$ and $b \in \mathbb{R}$, threshold $\delta \geq 0$

Output: approximate solution $\mathbf{w}$ and b

for $t = 1, 2, \ldots$ **do**
- receive index $I_t \in \{1, \ldots, n\}$ // I_t can be random
- **if** $\mathsf{y}_{I_t}(\langle \mathbf{x}_{I_t}, \mathbf{w} \rangle + b) \leq \delta$ **then**
  - $\mathbf{w} \leftarrow \mathbf{w} + \mathsf{y}_{I_t} \mathbf{x}_{I_t}$ // update after a "mistake"
  - $b \leftarrow b + \mathsf{y}_{I_t}$

- Stochastic gradient applied to $\ell_i(\mathbf{w}, b) = [\delta - \mathsf{y}_i(\langle \mathbf{x}_i, \mathbf{w} \rangle + b)]_+$
- Can now also employ a step size η_t in each step

Computing the Mean

$$f(\mathbf{w}) = \tfrac{1}{2}\mathbb{E}_{\mathbf{x}}\|\mathbf{w} - \mathbf{x}\|_2^2, \quad \ell_{\mathbf{x}} := \tfrac{1}{2}\|\mathbf{w} - \mathbf{x}\|_2^2$$

- Obviously, $\mathbf{w}_\star = \mathbb{E}[\mathbf{x}]$ is the mean, with $f_\star = \frac{1}{2}\mathbb{E}_{\mathbf{x}}\|\mathbf{x} - \mathbb{E}(\mathbf{x})\|_2^2$
- With stochastic gradient:
  - sample $\mathbf{x}_t$
  - compute $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t(\mathbf{w}_t - \mathbf{x}_t) = (1 - \eta_t)\mathbf{w}_t + \eta_t\mathbf{x}_t$
  - if $\mathbf{w}_0 = \mathbf{0}$ and $\eta_t = \frac{1}{t+1}$, then $\mathbf{w}_{t+1} = \frac{1}{t+1}\sum_{s=0}^{t} \mathbf{x}_s$
  - clearly, $\mathbf{w}_t \to \mathbb{E}[\mathbf{x}]$ and $\mathbb{E}f(\mathbf{w}_t) = (1 + \frac{1}{t})f_\star$
  - known to be statistically optimal for $d \leq 2$

Randomized Kaczmarz

$$f(\mathbf{w}) = \frac{1}{2n}\sum_{i=1}^{n}(\langle \mathbf{x}_i, \mathbf{w}\rangle - y_i)^2, \quad \ell_i := \tfrac{1}{2}(\langle \mathbf{x}_i, \mathbf{w}\rangle - y_i)^2$$

- For each i, $\|\mathbf{x}_i\|_2 = 1$ (w.l.o.g.)
- Assume there exists some $\mathbf{w}_\star$ so that $f(\mathbf{w}_\star) = 0$
- Can use constant step size $\eta_t \equiv 1$

$$\mathbf{w}_{t+1} = \mathbf{w}_t - (\langle \mathbf{x}_{i_t}, \mathbf{w}\rangle - y_{i_t})\mathbf{x}_{i_t} = \mathbf{w}_t - \mathbf{x}_{i_t}\mathbf{x}_{i_t}^\top(\mathbf{w}_t - \mathbf{w}_\star)$$

$$\mathbf{w}_{t+1} - \mathbf{w}_\star = \prod_{s=0}^{t}(I - \mathbf{x}_{i_s}\mathbf{x}_{i_s}^\top)(\mathbf{w}_s - \mathbf{w}_\star)$$

T. Strohmer and R. Vershynin. "A Randomized Kaczmarz Algorithm with Exponential Convergence". *Journal of Fourier Analysis and Applications*, vol. 15, no. 2 (2009), pp. 262–278.

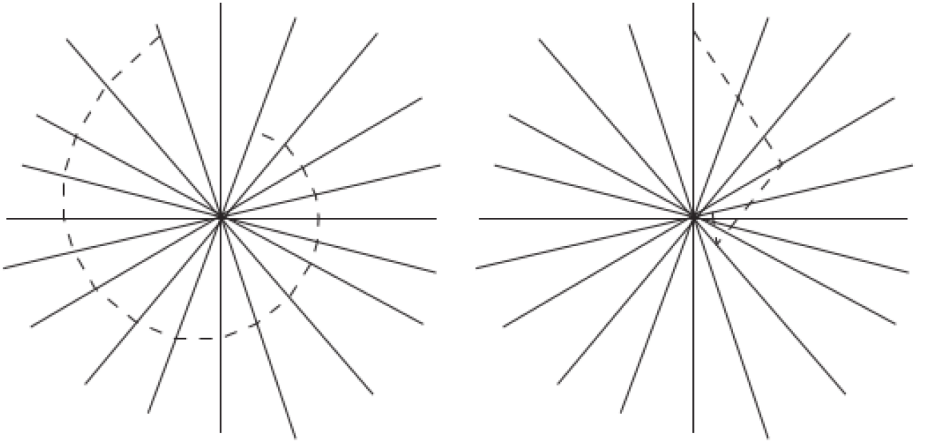

Figure 5 1 Kaczmarz method Deterministic, ordered choice (left) leads to slow convergence; randomized Kaczmarz (right) converges faster.

S. J. Wright and B. Recht. "Optimization for Data Analysis". Cambridge University Press, 2022.

Convergence Analysis

- Key assumption: controlled variance

$$\mathbb{E}\|\nabla f(\mathbf{w}, \boldsymbol{\xi})\|_2^2 \leq \mathsf{L}\|\mathbf{w} - \mathbf{w}_\star\|_2^2 + \sigma^2$$

- One step progress:

$$\begin{aligned}\|\mathbf{w}_{t+1} - \mathbf{w}_\star\|_2^2 &= \|\mathrm{P}_C[\mathbf{w}_t - \eta_t \nabla f(\mathbf{w}_t, \boldsymbol{\xi}_t)] - \mathrm{P}_C(\mathbf{w}_\star)\|_2^2 \\ &\leq \|\mathbf{w}_t - \eta_t \nabla f(\mathbf{w}_t, \boldsymbol{\xi}_t) - \mathbf{w}_\star\|_2^2 \\ &= \|\mathbf{w}_t - \mathbf{w}_\star\|_2^2 - 2\eta_t \langle \mathbf{w}_t - \mathbf{w}_\star, \nabla f(\mathbf{w}_t, \boldsymbol{\xi}_t)\rangle + \eta_t^2 \|\nabla f(\mathbf{w}_t, \boldsymbol{\xi}_t)\|_2^2\end{aligned}$$

- Conditioned on $\mathbf{w}_t$ and taking expectation w.r.t. $\boldsymbol{\xi}_t$:

$$\begin{aligned}\mathbb{E}_t[\|\mathbf{w}_{t+1} - \mathbf{w}_\star\|_2^2] &= \|\mathbf{w}_t - \mathbf{w}_\star\|_2^2 - 2\eta_t \langle \mathbf{w}_t - \mathbf{w}_\star, \nabla f(\mathbf{w}_t)\rangle + \eta_t^2 \mathbb{E}_t[\|\nabla f(\mathbf{w}_t, \boldsymbol{\xi}_t)\|_2^2] \\ &\leq \|\mathbf{w}_t - \mathbf{w}_\star\|_2^2 - 2\eta_t[f(\mathbf{w}_t) - f_\star] + \eta_t^2[\mathsf{L}\|\mathbf{w}_t - \mathbf{w}_\star\|_2^2 + \sigma^2] \\ &= (1 + \eta_t^2 \mathsf{L})\|\mathbf{w}_t - \mathbf{w}_\star\|_2^2 - 2\eta_t[f(\mathbf{w}_t) - f_\star] + \eta_t^2 \sigma^2\end{aligned}$$

$\mathsf{L} = 0$

$$\mathbb{E}[\|\mathbf{w}_{t+1} - \mathbf{w}_\star\|_2^2] \le \mathbb{E}[\|\mathbf{w}_t - \mathbf{w}_\star\|_2^2] - 2\eta_t \mathbb{E}[f(\mathbf{w}_t) - f_\star] + \eta_t^2\sigma^2$$

- Telescoping:

$$\sum_{s=0}^{t} 2\eta_s \mathbb{E}[f(\mathbf{w}_s) - f_\star] \le \mathbb{E}[\|\mathbf{w}_0 - \mathbf{w}_\star\|_2^2] + \sum_{s=0}^{t} \eta_s^2\sigma^2$$

- Defining $\mathbf{z}_t = \sum_{s=0}^t \eta_s \mathbf{w}_s / \sum_{s=0}^t \eta_s$ we obtain

$$\mathbb{E}[f(\mathbf{z}_t) - f_\star] \le \frac{\mathbb{E}[\|\mathbf{w}_0 - \mathbf{w}_\star\|_2^2] + \sum_{s=0}^t \eta_s^2\sigma^2}{2\sum_{s=0}^t \eta_s}$$

- Converges to 0 iff $\eta_t \to 0$ and $\sum_t \eta_t = \infty$
- With $\eta_t = O(1/\sqrt{t})$ we can obtain expected convergence rate $O(1/\sqrt{t})$

Logistic Regression

$$f(\mathbf{w}) := \frac{1}{n}\sum_{i=1}^{n} \ell_i(\mathbf{w}), \quad \text{where} \quad \ell_i(\mathbf{w}) = \log[1 + \exp(-\mathsf{y}_i \langle \mathbf{x}_i, \mathbf{w} \rangle)]$$

- We clearly have

$$\nabla \ell_i(\mathbf{w}) = -\frac{\exp(-\mathsf{y}_i \langle \mathbf{x}_i, \mathbf{w} \rangle)}{1 + \exp(-\mathsf{y}_i \langle \mathbf{x}_i, \mathbf{w} \rangle)} \mathsf{y}_i \mathbf{x}_i$$

- Can choose $\mathsf{L} = 0$ and

$$\sigma^2 = \max_i \|\mathbf{x}_i\|_2^2$$

$\sigma = 0$

$$\mathbb{E}_t[\|\mathbf{w}_{t+1} - \mathbf{w}_\star\|_2^2] \leq (1 + \eta_t^2 \mathsf{L})\|\mathbf{w}_t - \mathbf{w}_\star\|_2^2 - 2\eta_t[f(\mathbf{w}_t) - f_\star]$$

- Assume further that f is μ-strongly convex:

$$f(\mathbf{w}) - f_\star \geq \frac{\mu}{2}\|\mathbf{w} - \mathbf{w}_\star\|_2^2$$

- Thus, we have the recursion:

$$\mathbb{E}[\|\mathbf{w}_{t+1} - \mathbf{w}_\star\|_2^2] \leq (1 - \eta_t\mu + \eta_t^2 \mathsf{L})\mathbb{E}[\|\mathbf{w}_t - \mathbf{w}_\star\|_2^2]$$

- Linear (expected) convergence if $\eta_t \in (0, \frac{\mu}{\mathsf{L}})$, with optimal $\eta = \frac{\mu}{2\mathsf{L}}$ such that

$$\mathbb{E}[\|\mathbf{w}_{t+1} - \mathbf{w}_\star\|_2^2] \leq (1 - \frac{\mu^2}{4\mathsf{L}})\mathbb{E}[\|\mathbf{w}_t - \mathbf{w}_\star\|_2^2]$$

Randomized Kaczmarz

$$f(\mathbf{w}) := \frac{1}{n}\sum_{i=1}^{n} \ell_i(\mathbf{w}), \quad \text{where} \quad \ell_i(\mathbf{w}) = \tfrac{1}{2}[y_i - \langle \mathbf{x}_i, \mathbf{w}\rangle]^2$$

- Assuming $f(\mathbf{w}_\star) = 0$, we have

$$\nabla \ell_i(\mathbf{w}) = (\langle \mathbf{x}_i, \mathbf{w}\rangle - y_i)\mathbf{x}_i = \mathbf{x}_i\mathbf{x}_i^\top(\mathbf{w} - \mathbf{w}_\star)$$

- Can choose $\sigma = 0$ and

$$\mathbb{E}\|\mathbf{x}_i\mathbf{x}_i^\top(\mathbf{w} - \mathbf{w}_\star)\|_2^2 \le \underbrace{\mathbb{E}\|\mathbf{x}_i\|_2^4}_{\mathsf{L}} \cdot \|\mathbf{w} - \mathbf{w}_\star\|_2^2$$

- f is indeed strongly convex: $\nabla^2 f(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^\top$

General Case

$$\mathbb{E}_t[\|\mathbf{w}_{t+1} - \mathbf{w}_\star\|_2^2] \le (1 + \eta_t^2 \mathsf{L})\|\mathbf{w}_t - \mathbf{w}_\star\|_2^2 - 2\eta_t[f(\mathbf{w}_t) - f_\star] + \eta_t^2\sigma^2$$

- Assume further that f is μ-strongly convex:

$$f(\mathbf{w}) - f_\star \ge \frac{\mu}{2}\|\mathbf{w} - \mathbf{w}_\star\|_2^2$$

- Telescoping: $\mathbb{E}[\|\mathbf{w}_{t+1} - \mathbf{w}_\star\|_2^2] \le (1 - \eta_t\mu + \eta_t^2\mathsf{L})\mathbb{E}[\|\mathbf{w}_t - \mathbf{w}_\star\|_2^2] + \eta_t^2\sigma^2$

$$\mathbb{E}[\|\mathbf{w}_{t+1} - \mathbf{w}_\star\|_2^2] \le \prod_{s=0}^{t}(1 - \eta_s\mu + \eta_s^2\mathsf{L}) \cdot [\|\mathbf{w}_0 - \mathbf{w}_\star\|_2^2] + \sum_{k=0}^{t}\eta_k^2\sigma^2 \prod_{s=k+1}^{t}(1 - \eta_s\mu + \eta_s^2\mathsf{L})$$

- With $\eta_t = \frac{1}{2\mathsf{L}^2/\sigma + \sigma t}$ we have $\mathbb{E}[\|\mathbf{w}_{t+1} - \mathbf{w}_\star\|_2^2] = O(\eta_t)$

Minibatching

- Some people consider the variance instead:

$$\mathbb{E}\|\nabla f(\mathbf{w}, \boldsymbol{\xi}) - \nabla f(\mathbf{w})\|_2^2 = \mathbb{E}\|\nabla f(\mathbf{w}, \boldsymbol{\xi})\|_2^2 - \|\nabla f(\mathbf{w})\|_2^2$$

- If averaging the stochastic gradient over a minibatch of size b:

$$\mathbf{g} = \frac{1}{b}\sum_{k=1}^{b} \nabla f(\mathbf{w}, \boldsymbol{\xi}_{i_k})$$

  - increase computation by a factor of b
  - decrease variance by a factor of b too
  - suitable for parallel computation

Incremental Gradient

$$f(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n} \ell_i(\mathbf{w})$$

Let $\mathbf{w}_i^* \in \operatorname{argmin} \ell_i$, where ℓ_i is L_i-smooth. Then,

$$\begin{aligned}
\mathbb{E}_I \|\nabla \ell_I(\mathbf{w})\|_2^2 &\le \mathbb{E}\left[\mathsf{L}_I^2 \|\mathbf{w} - \mathbf{w}_I^*\|_2^2\right] \\
&\le \mathbb{E}\left[2\mathsf{L}_I^2 (\|\mathbf{w} - \mathbf{w}_\star\|_2^2 + \|\mathbf{w}_i^* - \mathbf{w}_\star\|_2^2)\right] \\
&= \underbrace{\frac{2}{n}\sum_{i=1}^{n} \mathsf{L}_i^2}_{\mathsf{L}} \cdot \|\mathbf{w} - \mathbf{w}_\star\|_2^2 + \underbrace{\frac{2}{n}\sum_{i=1}^{n} \mathsf{L}_i^2 \|\mathbf{w}_i^* - \mathbf{w}_\star\|_2^2}_{\sigma^2}
\end{aligned}$$