Large-scale Machine Learning and Optimization

Zaiwen Wen

http://bicmr.pku.edu.cn/~wenzw/bigdata2020.html

Acknowledgement: this slides is based on Dimitris Papailiopoulos and Shiqian Ma lecture notes and the book "understanding machine learning theory and algorithms" of Shai Shalev-Shwartz and Shai Ben-David.

Thanks Yongfeng Li for preparing part of this slides

Why Optimization in Machine Learning?

Many problems in ML can be written as

$$\min_{\theta \in \mathcal{W}} \quad \sum_{i=1}^{N} \frac{1}{2} \|x_i^\top \theta - y_i\|_2^2 + \mu \|\theta\|_2^2 \quad \text{linear regression}$$

$$\min_{\theta \in \mathcal{W}} \quad \frac{1}{N} \sum_{i=1}^{N} \log(1 + \exp(-y_i x_i^\top \theta)) + \mu \|\theta\|_2^2 \quad \text{logistic regression}$$

$$\min_{\theta \in \mathcal{W}} \quad \sum_{i=1}^{N} \ell(h(\theta, x_i), y_i) + \mu \varphi(\theta) \quad \text{general formulation}$$

- The pairs (x_i, y_i) are given data, y_i is the label of the data point x_i
- $\ell(\cdot)$: measures how model fit for data points (avoids under-fitting)
- $\varphi(\theta)$: regularization term (avoids over-fitting)
- $h(\theta, x)$: linear function or models constructed from deep neural networks

Sparse Logistic Regression

The logistic regression problem:

$$\min_{\theta \in \mathbb{R}^n} \quad \frac{1}{N} \sum_{i=1}^{N} \log(1 + \exp(-y_i x_i^T \theta)) + \mu \|\theta\|_2^2.$$

- The data pair $\{x_i, y_i\} \in \mathbb{R}^n \times \{-1, 1\}, i \in [N]$,

Data Set	# data N	# features n	sparsity(%)
cina	16,033	132	70.49
a9a	32,561	123	88.72
ijcnn1	49,990	22	40.91
covtype	581,012	54	77.88
url	2,396,130	3,231,961	99.99
susy	5,000,000	18	1.18
higgs	11,000,000	28	7.89
news20	19,996	1,355,191	99.97
rcv1	20,242	47,236	99.84
kdda	8,407,752	20,216,830	99.99

Deep Learning

The objective function is the CrossEntropy function plus regularization term:

$$\min_{\theta} \quad \frac{1}{N}\sum_{i=1}^{N} -\log\left(\frac{\exp(h(\theta,x_i)[y_i])}{\sum_j \exp(h(\theta,x_i)[y_j])}\right) + \mu\|\theta\|_2^2$$

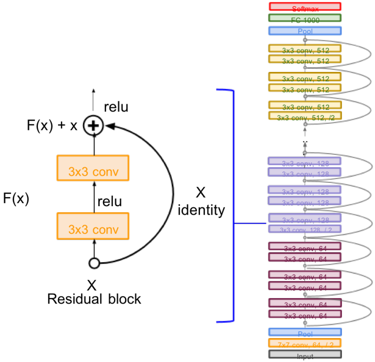
where $h(\theta, x_i)$ is output from network, and (x_i, y_i) are data points.

	Cifar-10	Cifar-100
# num_class	10	100
# number per class (training set)	5,000	500
# number per class (testing set)	1,000	100
# Total parametes of VGG-16	15,253,578	15,299,748
# Total parameters of ResNet-18	11,173,962	11,220,132

Table: A description of datasets used in the neural network experiments

ResNet Architecture

- Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, Cited by **114474 since 2015** at Google scholar
- Stack residual blocks. Every residual block has two 3x3 conv layers.
- Make networks from shallow to deep.
- Fancy network architecture. Many Applications.
- High-computationally-cost !
- ResNet-50 on ImageNet, **29 hours using 8 Tesla P100 GPUs**



Outline

Machine Learning Model

Machine learning model:

- $(x, y) \sim \mathcal{P}$, $\mathcal{P}$ is a underlying distribution.
- Given a dataset $\mathcal{D} = \{(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)\}$. $(x_i, y_i) \sim \mathcal{P}$ i.i.d.
- Our goal is to find a hypothesis $h(\theta, x)$ with the smallest expected risk, i.e.,

$$\min_{h \in \mathcal{H}} R[h] := \mathbf{E}[\ell(h(\theta, x), y)] \tag{1}$$

  where $\mathcal{H}$ is hypothesis class.

Machine Learning Model

- In practice, we don not know the exact form of the underlying distribution $\mathcal{P}$.
- Empirical Risk Minimization (ERM)

$$\min_{h \in \mathcal{H}} \hat{R}_n[h] := \frac{1}{n} \sum_{i=1}^{n} \ell(h(\theta, x_i), y_i) \tag{2}$$

- We care about two questions on ERM:
  - When does the ERM concentrate around the true risk?
  - How does the hypothesis class affect the ERM?

Machine Learning Model

- Empirical risk minimizer $\hat{h}_n^* \in \underset{h \in \mathcal{H}}{\operatorname{argmin}} \hat{R}_n[h]$
- Expected risk minimizor $h^* \in \underset{h \in \mathcal{H}}{\operatorname{argmin}} R[h]$
- The concentration means that for any $\epsilon > 0, 0 < \delta < 1$, if n is larger enough, we have

$$\mathcal{P}(|R[\hat{h}_n^*] - R[h^*]| \leq \epsilon) > 1 - \delta \tag{3}$$

- It just means that $R[\hat{h}_n^*]$ convergences to $R[h^*]$ in probability.
- The concentration will fail in some cases

Hoeffding Inequality

Let $X_1, X_2, \cdots$ be a sequence of i.i.d. random variables and assume that for all i, $E(X_i) = \mu$ and $\mathcal{P}(a \leq X_i \leq b) = 1$. Then for any $\epsilon > 0$

$$\mathcal{P}\left(|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu| \geq \epsilon\right) \leq 2\exp\left(-\frac{2n\epsilon^2}{(b-a)^2}\right) \tag{4}$$

- The Hoeffding Inequality describes the asymptotic property that sampling mean convergences to expectation.
- Azuma-Hoeffding inequality is a martingle version. Let $X_1, X_2, \cdots$ be a martingale difference sequence with $|X_i| \leq B$ for all $i = 1, 2, ...$ Then

$$\mathcal{P}(\sum_{i=1}^{n} X_i \geq t) \leq \exp\left(-\frac{2t^2}{nB^2}\right) \tag{5}$$

$$\mathcal{P}(\sum_{i=1}^{n} X_i \leq t) \leq \exp\left(-\frac{2t^2}{nB^2}\right) \tag{6}$$

- To make the exposition simpler, we assume that our loss function, $0 \leq \ell(a,b) \leq 1, \forall a, b$.
- By Hoeffding Inequality, fixed h

$$\mathcal{P}(|\hat{R}_n[h] - R[h]| \geq \epsilon) \leq 2e^{-2n\epsilon^2} \tag{7}$$

- Union Bound

$$\mathcal{P}(\bigcup_{h \in \mathcal{H}} \{|\hat{R}_n[h] - R[h]| \geq \epsilon\}) \leq 2|\mathcal{H}|e^{-2n\epsilon^2} \tag{8}$$

- If we want to bound $\mathcal{P}(\bigcup_{h \in \mathcal{H}} \{|\hat{R}_n[h] - R[h]| \geq \epsilon\}) \leq \delta$, we need the size of sample

$$n \geq \frac{1}{2\epsilon^2} \log\left(\frac{2|\mathcal{H}|}{\delta}\right) = O\left(\frac{\log|\mathcal{H}| + \log(\delta^{-1})}{\epsilon^2}\right) \tag{9}$$

- What if $|\mathcal{H}| = \infty$? This bound doesn't work

- If n is large enough, with a probability $1-\delta$, we have

$$\begin{aligned} R[\hat{h}_n^*] - R[h^*] &= (R[\hat{h}_n^*] - \hat{R}_n[\hat{h}_n^*]) + (\hat{R}_n[\hat{h}_n^*] - \hat{R}_n[h^*]) \\ &\quad + (\hat{R}_n[h^*] - R[h^*]). \\ &\leq \epsilon + 0 + \epsilon. \end{aligned}$$

- For a two label classification problem, with a probability $1-\delta$, we have

$$\sup_{h \in \mathcal{H}} |\hat{R}_n[h] - R[h]| \leq O\left(\sqrt{\frac{VC[\mathcal{H}]\log(\frac{n}{VC[\mathcal{H}]}) + \log(\frac{1}{\delta}))}{n}}\right) \quad (10)$$

  where $VC[\mathcal{H}]$ is a VC dimension of $\mathcal{H}$.
- Finite VC dimension is sufficient and necessary condition of empirical risk concentration for two label classification.

VC dimension

- VC dimension of a set-family: Let H be a set family (a set of sets) and C a set. Their intersection is defined as the following set-family:
$$H \cap C := \{h \cap C \mid h \in H\}$$
We say that a set C is shattered by H if $H \cap C = 2^C$.
The **VC dimension** of H is the largest integer D such that there exists a set C with cardinality D that is shattered by H.
- A classification model f with some parameter vector θ is said to shatter a set of data points $(x_1, x_2, \ldots, x_n)$ if, for all assignments of labels to those points, there exists a θ such that the model f makes no errors when evaluating that set of data points.
The **VC dimension** of a model f is the maximum number of points that can be arranged so that f shatters them. More formally, it is the maximum cardinal D such that some data point set of cardinality D can be shattered by f.

VC dimension

- example:
  - If $\forall n$ and $\{(x_1, y_1), \cdots, (x_n, y_n)\}$, there exits $h \in \mathcal{H}$ s.t. $h(x_i) = y_i$, then $VC[\mathcal{H}] = \infty$
  - For a neural network whose activation functions are all sign functions, then $VC[\mathcal{H}] \leq O(w \log(w))$, where w is the number of parameters.
- We must use prior knowledge and choose a proper hypothesis class.
- Suppose a is the true model

$$R[\hat{h}_n^*] - R[a] = \underbrace{(R[\hat{h}_n^*] - R[h^*])}_{A} + \underbrace{(R[h^*] - R[a])}_{B}$$

- If the hypothesis class is too large, B will be small but A will be large. (overfitting)
- If the hypothesis class is too small, A will be small but B will be large. (underfitting)

Outline

The gradient and subgradient methods

- Consider the problem

$$\min_{x \in \mathbb{R}^n} f(x) \tag{11}$$

- **gradient methods**

$$x_{k+1} = x_k - \alpha_k \nabla f(x_k) \tag{12}$$

- **subgradient methods**

$$x_{k+1} = x_k - \alpha_k g_k, g_k \in \partial f(x_k) \tag{13}$$

- the update is equal to

$$x_{k+1} = \arg\min_x f(x_k) + \langle g_k, x - x_k \rangle + \frac{1}{2\alpha_k} ||x - x_k||_2^2 \tag{14}$$

Convergence guarantees

Assumption

- There is at least one minimizing point $x^* \in \arg\min_x f(x)$ with $f(x^*) > -\infty$
- The subgradients are bounded: $||g||_2 \leq M \leq \infty$ for all x and all $g \in \partial f(x)$.

Theorem 1: Convergence of subgradient

Let $\alpha_k \geq 0$ be any non-negative sequence of stepsizes and the preceding assumptions hold. Let x_k be generated by the subgradient iteration. Then for all $K \geq 1$,

$$\sum_{k=1}^{K} \alpha_k[f(x_k) - f(x^*)] \leq \frac{1}{2}||x_1 - x^*||_2^2 + \frac{1}{2}\sum_{k=1}^{K} \alpha_k^2 M^2. \quad (15)$$

Proof of Theorem 1

- By convexity
$$\langle g_k, x^* - x_k \rangle \le f(x^*) - f(x_k)$$

-
$$\begin{aligned} & \frac{1}{2}||x_{k+1} - x^*||_2^2 = \frac{1}{2}||x_k - \alpha_k g_k - x^*||_2^2 \\ = \; & \frac{1}{2}||x_k - x^*||_2^2 + \alpha_k \langle g_k, x^* - x_k \rangle + \frac{\alpha_k^2}{2}||g_2||_2^2 \\ \le \; & \frac{1}{2}||x_k - x^*||_2^2 - \alpha_k(f(x_k) - f(x^*)) + \frac{\alpha_k^2}{2}M^2 \end{aligned}$$

-
$$\alpha_k(f(x_k) - f(x^*)) \le \frac{1}{2}||x_k - x^*||_2^2 - \frac{1}{2}||x_{k+1} - x^*||_2^2 + \frac{\alpha_k^2}{2}M^2$$

Convergence guarantees

Corollary

Let $A_k = \sum_{i=1}^{K} \alpha_i$ and define $\bar{x}_K = \frac{1}{A_K} \sum_{k=1}^{K} \alpha_k x_k$

$$f(\bar{x}_k) - f(x^*) \leq \frac{||x_1 - x^*||_2^2 + \sum_{k=1}^{K} \alpha_k^2 M^2}{2 \sum_{k=1}^{K} \alpha_k}. \quad (16)$$

- Convergence: $\sum_{k=1}^{\infty} \alpha_k = \infty$ and $\frac{\sum_{k=1}^{K} \alpha_k^2}{\sum_{k=1}^{K} \alpha_k} \to 0$
- Let $||x_1 - x^*|| \leq R$. For a fixed stepsize $\alpha_k = \alpha$:

$$f(\bar{x}_K) - f(x^*) \leq \frac{R^2}{2K\alpha} + \frac{\alpha M^2}{2}$$

- For a given K, take $\alpha = \frac{R}{M\sqrt{K}}$:

$$f(\bar{x}_K) - f(x^*) \leq \frac{RM}{\sqrt{K}}$$

Projected subgradient methods

- Consider the problem

$$\min_{x \in C} f(x) \tag{17}$$

- **subgradient methods**

$$x_{k+1} = \pi_C(x_k - \alpha_k g_k), g_k \in \partial f(x_k) \tag{18}$$

- projection: $\pi_C(x) = \arg\min_{y \in C} ||x - y||_2^2$
- the update is equal to

$$x_{k+1} = \arg\min_{x \in C} f(x_k) + \langle g_k, x - x_k \rangle + \frac{1}{2\alpha_k}||x - x_k||_2^2 \tag{19}$$

Convergence guarantees

Assumption

- The set $C \subset \mathbb{R}^n$ is compact and convex, and $||x - x^*||_2 \leq R < \infty$ for all $x \in C$.
- The subgradients are bounded: $||g||_2 \leq M \leq \infty$ for all x and all $g \in \partial f(x)$.

Theorem 2: Convergence of projected subgradient method

Let $\alpha_k \geq 0$ be any non-negative sequence of stepsizes and the preceding assumptions hold. Let x_k be generated by the projected subgradient iteration. Then for all $K \geq 1$,

$$\sum_{k=1}^{K} \alpha_k [f(x_k) - f(x^*)] \leq \frac{1}{2} ||x_1 - x^*||_2^2 + \frac{1}{2} \sum_{k=1}^{K} \alpha_k^2 M^2. \tag{20}$$

Proof of Theorem 2

- By non-expansiveness of $\pi_C(x)$

$$\|x_{k+1} - x^*\|_2^2 = \|\pi_C(x_k - \alpha g_k) - x^*\| \leq \|x_k - \alpha g_k - x^*\|$$

-

$$\begin{aligned} \frac{1}{2}||x_{k+1} - x^*||_2^2 &\leq \frac{1}{2}||x_k - \alpha_k g_k - x^*||_2^2 \\ &= \frac{1}{2}||x_k - x^*||_2^2 + \alpha_k \langle g_k, x^* - x_k \rangle + \frac{\alpha_k^2}{2}||g_2||_2^2 \\ &\leq \frac{1}{2}||x_k - x^*||_2^2 - \alpha_k(f(x_k) - f(x^*)) + \frac{\alpha_k^2}{2}M^2 \end{aligned}$$

-

$$\alpha_k(f(x_k) - f(x^*)) \leq \frac{1}{2}||x_k - x^*||_2^2 - \frac{1}{2}||x_{k+1} - x^*||_2^2 + \frac{\alpha_k^2}{2}M^2$$

Convergence guarantees

Corollary

Let $A_k = \sum_{i=1}^{k} \alpha_i$ and define $\bar{x}_K = \frac{1}{A_K} \sum_{k=1}^{K} \alpha_k x_k$

$$f(\bar{x}_K) - f(x^*) \leq \frac{||x_1 - x^*||_2^2 + \sum_{k=1}^{K} \alpha_k^2 M^2}{2 \sum_{k=1}^{K} \alpha_k}. \quad (21)$$

- Convergence: $\sum_{k=1}^{\infty} \alpha_k = \infty$ and $\frac{\sum_{k=1}^{K} \alpha_k^2}{\sum_{k=1}^{K} \alpha_k} \to 0$
- a fixed stepsize $\alpha_k = \alpha$:

$$f(\bar{x}_K) - f(x^*) \leq \frac{R^2}{2K\alpha} + \frac{\alpha M^2}{2}$$

- Take $\alpha = \frac{R}{M\sqrt{K}}$:

$$f(\bar{x}_K) - f(x^*) \leq \frac{RM}{\sqrt{K}}$$

Stochastic subgradient methods

- the stochastic optimization problem

$$\min_{x \in C} f(x) := \mathbf{E}_{\mathrm{P}}[F(x;S)] \tag{22}$$

- S is a random space is a random variable on the space $\mathcal{S}$ with distribution P.
- for each s, $x \to F(x;s)$ is convex.
- The subgradient $\mathbf{E}_{\mathrm{P}}[g(x;S)] \in \partial f(x)$, where $g(x;s) \in \partial F(x;s)$.
-

$$\begin{aligned} f(y) &= \mathbf{E}_{\mathrm{P}}[F(y;S)] \geq \mathbf{E}_{\mathrm{P}}[F(x;S) + \langle g(x,S), y-x \rangle] \\ &= f(x) + \langle \mathbf{E}_{\mathrm{P}}[g(x;S)], y-x \rangle \end{aligned}$$

Stochastic subgradient methods

- the deterministic optimization problem

$$\min_{x\in C} f(x) := \frac{1}{m}\sum_{i=1}^{m} F(x; s_i) \tag{23}$$

- Why Stochastic?
  - $\mathbf{E}_{\mathrm{P}}[F(x; S)]$ is generally intracktable to compute
  - Small complexity: only one subgradient $g(x; s) \in \partial F(x; s)$ needs to be computed in one iteration.
  - More possible to get global solution for non-convex case.
- stochasitic subgradient method

$$x_{k+1} = \pi_C(x_k - \alpha_k g_k), \ \mathbf{E}[g_k | x_k] \in \partial f(x_k)$$

Convergence guarantees

Assumption

- The set $C \subset \mathbb{R}^n$ is compact and convex, and $||x - x^*||_2 \leq R < \infty$ for all $x \in C$.
- The variance are bounded: $\mathbf{E}||g(x,S)||_2^2 \leq M^2 \leq \infty$ for all x.

Theorem 3: Convergence of stochastic subgradient method

Let $\alpha_k \geq 0$ be any non-negative sequence of stepsizes and the preceding assumptions hold. Let x_k be generated by the stochastic subgradient iteration. Then for all $K \geq 1$,

$$\sum_{k+1}^{K} \alpha_k \mathbf{E}(f(x_k) - f(x^*)) \leq \frac{1}{2}\mathbf{E}||x_1 - x^*||_2^2 + \frac{\sum_{k=1}^{K} \alpha_k^2 M^2}{2} \tag{24}$$

Proof of Theorem 2

- Let $f'(x_k) = E[g_k|x_k]$ and $\xi_k = g_k - f'(x_k)$,

$$\begin{aligned} & \frac{1}{2}||x_{k+1} - x^*||_2^2 \leq \frac{1}{2}||x_k - \alpha_k g_k - x^*||_2^2 \\ = & \frac{1}{2}||x_k - x^*||_2^2 + \alpha_k \langle g_k, x^* - x_k \rangle + \frac{\alpha_k^2}{2}||g_k||_2^2 \\ = & \frac{1}{2}||x_k - x^*||_2^2 + \alpha_k \left\langle f'(x_k), x^* - x_k \right\rangle + \frac{\alpha_k^2}{2}||g_k||_2^2 + \alpha_k \langle \xi_k, x^* - x_k \rangle \\ \leq & \frac{1}{2}||x_k - x^*||_2^2 - \alpha_k(f(x_k) - f(x^*)) + \frac{\alpha_k^2}{2}||g_k||_2^2 + \alpha_k \langle \xi_k, x^* - x_k \rangle \end{aligned}$$

- $$\mathbf{E}[\langle \xi_k, x^* - x_k \rangle] = \mathbf{E}[\mathbf{E}[\langle \xi_k, x^* - x_k \rangle \,|x_k]] = 0.$$

- $$\alpha_k \mathbf{E}(f(x_k) - f(x^*)) \leq \frac{1}{2}\mathbf{E}||x_k - x^*||_2^2 - \frac{1}{2}\mathbf{E}||x_{k+1} - x^*||_2^2 + \frac{\alpha_k^2}{2}M^2$$

Convergence guarantees

Corollary

Let $A_k = \sum_{i=1}^{k} \alpha_i$ and define $\bar{x}_K = \frac{1}{A_K} \sum_{k=1}^{K} \alpha_k x_k$

$$\mathbf{E}[f(\bar{x}_K) - f(x^*)] \le \frac{R^2 + \sum_{k=1}^{K} \alpha_k^2 M^2}{2 \sum_{k=1}^{K} \alpha_k}. \tag{25}$$

- Convergence: $\sum_{k=1}^{\infty} \alpha_k = \infty$ and $\frac{\sum_{k=1}^{K} \alpha_k^2}{\sum_{k=1}^{K} \alpha_k} \to 0$
- a fixed stepsize $\alpha_k = \alpha$:

$$\mathbf{E}(f(\bar{x}_K) - f(x^*)) \le \frac{R^2}{2K\alpha} + \frac{\alpha M^2}{2}$$

- Take $\alpha = \frac{R}{M\sqrt{K}}$:

$$\mathbf{E}(f(\bar{x}_K) - f(x^*)) \le \frac{RM}{\sqrt{K}}$$

Theorem 5: Convergence of stochastic subgradient method

Let $\alpha_k > 0$ be non-increasing sequence of stepsizes and the preceding assumptions hold. Let $\bar{x} = \frac{1}{K}\sum_{k=1}^{K} x_k$. Then,

$$\mathbf{E}[f(\bar{x}_K) - f(x^*)] \leq \frac{R^2}{2K\alpha_K} + \frac{1}{2K}\sum_{k=1}^{K} \alpha_k M^2. \quad (26)$$

- $$\alpha_k \mathbf{E}(f(x_k) - f(x^*)) \leq \frac{1}{2}\mathbf{E}||x_k - x^*||_2^2 - \frac{1}{2}\mathbf{E}||x_{k+1} - x^*||_2^2 + \frac{\alpha_k^2}{2}M^2$$

- $$\mathbf{E}(f(x_k) - f(x^*)) \leq \frac{1}{2\alpha_k}\mathbf{E}||x_k - x^*||_2^2 - \frac{1}{2\alpha_k}\mathbf{E}||x_{k+1} - x^*||_2^2 + \frac{\alpha_k}{2}M^2$$

Corollary

Let the conditions of Theorem 5 hold, and let $\alpha_k = \frac{R}{M\sqrt{k}}$ for each k. Then,

$$\mathbf{E}[f(\bar{x}_K) - f(x^*)] \leq \frac{3RM}{2\sqrt{K}}. \tag{27}$$

- proof

$$\sum_{k=1}^{K} \frac{1}{\sqrt{k}} \leq \int_0^K \frac{1}{\sqrt{t}} dt = 2\sqrt{K}.$$

Corollary

Let α_k be chosen such that $\mathbf{E}[f(\bar{x}_K) - f(x^*)] \to 0$. Then $f(\bar{x}_K) - f(x^*) \xrightarrow{\text{P}} 0$ as $K \to \infty$, that is, for all $\epsilon > 0$ we have

$$\limsup_{k\to\infty} \text{P}(f(\bar{x}_K) - f(x^*) \geq \epsilon) = 0. \tag{28}$$

- By markov inequality: $P(X \geq \alpha) \leq \frac{EX}{\alpha}$ if $X \geq 0$ and $\alpha > 0$

$$\mathrm{P}(f(\bar{x}_K) - f(x^*) \geq \epsilon) \leq \frac{1}{\epsilon}\mathbf{E}[f(\bar{x}_K) - f(x^*)] \to 0$$

Theorem 6: Convergence of stochastic subgradient method

In addition to the conditions of Theorem 5, assume that $||g||_2 \leq M$ for all stochastic subgradients g, Then for anything $\epsilon > 0$,

$$f(\bar{x}_K) - f(x^*) \leq \frac{R^2}{2K\alpha_K} + \frac{1}{2K}\sum_{k=1}^{K}\alpha_k M^2 + \frac{RM}{\sqrt{K}}\epsilon. \qquad (29)$$

with probability at least $1 - e^{-\frac{1}{2}\epsilon^2}$

- Taking $\alpha_k = \frac{R}{\sqrt{k}M}$ and setting $\delta = e^{-\frac{1}{2}\epsilon^2}$

$$f(\bar{x}_K) - f(x^*) \leq \frac{3RM}{2\sqrt{K}} + \frac{RM\sqrt{2\log\frac{1}{\delta}}}{\sqrt{K}}.$$

with probability at least $1 - \delta$

Azuma-Hoeffding Inequality

- **martingle**: A sequence $X_1, X_2, \cdots$ of random vectors is a martingale if there is a sequence of random vectors $Z_1, Z_2, \cdots$ such that for each n,
  - X_n is a function of Z_n,
  - Z_{n-1} is a function of Z_n,
  - we have the conditional expectation condition

$$\mathbf{E}[X_n | Z_{n-1}] = X_{n-1}.$$

- **martingale difference sequence** $X_1, X_2, \cdots$ is a martingale difference sequence if $S_n = \sum_{i=1}^n X_i$ is a martingle or, equivalently

$$\mathbf{E}[X_n | Z_{n-1}] = 0.$$

- **example** $X_1, X_2, \cdots$ independent and $E(X_i) = 0$, $Z_i = (X_1, \cdots, X_i)$.

Azuma-Hoeffding Inequality

Let $X_1, X_2, \cdots$ be a martingale difference sequence with $|X_i| \leq B$ for all $i = 1, 2, \ldots$ Then

$$\mathrm{P}(\sum_{i=1}^{n} X_i \geq t) \leq \exp(-\frac{2t^2}{nB^2}) \tag{30}$$

$$\mathrm{P}(\sum_{i=1}^{n} X_i \leq t) \leq \exp(-\frac{2t^2}{nB^2}) \tag{31}$$

- Let $\delta = \frac{t}{n}$

$$\mathrm{P}(\frac{1}{n}\sum_{i=1}^{m} X_i \geq \delta) \leq \exp(-\frac{2n\delta^2}{B^2})$$

- $X_1, X_2, \cdots$ i.i.d, $EX_i = \mu$

$$\mathrm{P}(|\frac{1}{n}\sum_{i=1}^{m} X_i - \mu| \geq \delta) \leq 2\exp(-\frac{2n\delta^2}{B^2})$$

Azuma-Hoeffding Inequality

Theorem 5: Convergence of stochastic subgradient method

Let $\alpha_k > 0$ be non-increasing sequence of stepsizes and the preceding assumptions hold. Let $\bar{x} = \frac{1}{K}\sum_{k=1}^{K} x_k$. Then,

$$\mathbf{E}[f(\bar{x}_K) - f(x^*)] \leq \frac{R^2}{2K\alpha_K} + \frac{1}{2K}\sum_{k=1}^{K}\alpha_k M^2. \quad (32)$$

Theorem 6: Convergence of stochastic subgradient method

In addition to the conditions of Theorem 5, assume that $||g||_2 \leq M$ for all stochastic subgradients g, Then for anything $\epsilon > 0$,

$$f(\bar{x}_K) - f(x^*) \leq \frac{R^2}{2K\alpha_K} + \frac{1}{2K}\sum_{k=1}^{K}\alpha_k M^2 + \frac{RM}{\sqrt{K}}\epsilon. \quad (33)$$

with probability at least $1 - e^{-\frac{1}{2}\epsilon^2}$

Proof of Theorem 6

- Let $f'(x_k) = E[g_k|x_k]$ and $\xi_k = g_k - f'(x_k)$,

$$
\begin{aligned}
& \frac{1}{2}||x_{k+1} - x^*||_2^2 = \frac{1}{2}||x_k - \alpha_k g_k - x^*||_2^2 \\
= \quad & \frac{1}{2}||x_k - x^*||_2^2 + \alpha_k \langle g_k, x^* - x_k \rangle + \frac{\alpha_k^2}{2}||g_k||_2^2 \\
= \quad & \frac{1}{2}||x_k - x^*||_2^2 + \alpha_k \left\langle f'(x_k), x^* - x_k \right\rangle + \frac{\alpha_k^2}{2}||g_k||_2^2 + \alpha_k \langle \xi_k, x^* - x_k \rangle \\
\leq \quad & \frac{1}{2}||x_k - x^*||_2^2 - \alpha_k(f(x_k) - f(x^*)) + \frac{\alpha_k^2}{2}||g_k||_2^2 + \alpha_k \langle \xi_k, x^* - x_k \rangle
\end{aligned}
$$

- $$f(x_k) - f(x^*) \leq \frac{1}{2\alpha_k}||x_k - x^*||_2^2 - \frac{1}{2\alpha_k}||x_{k+1} - x^*||_2^2 + \frac{\alpha_k}{2}||g_k||_2^2 + \langle \xi_k, x^* - x_k \rangle$$

Proof of Theorem 6

-

$$f(\bar{x}_K) - f(x) \le \frac{1}{K}\sum_{k=1}^{K} f(x_k) - f(x^*)$$

$$\le \quad \frac{1}{2K\alpha_K}||x_1 - x^*||_2^2 + \frac{1}{2K}\sum_{k=1}^{K}\alpha_k||g_k||_2^2 + \frac{1}{K}\sum_{i=1}^{K}\langle \xi_k, x^* - x_k\rangle$$

$$\le \quad \frac{1}{2K\alpha_K}||x_1 - x^*||_2^2 + \frac{1}{2K}\sum_{k=1}^{K}\alpha_k M^2 + \frac{1}{K}\sum_{i=1}^{K}\langle \xi_k, x^* - x_k\rangle$$

- Let $\omega = \frac{1}{2K\alpha_K}||x_1 - x^*||_2^2 + \frac{1}{2K}\sum_{k=1}^{K}\alpha_k M^2$

$$\mathrm{P}(f(\bar{x}_K) - f(x) - \omega \ge t) \le \mathrm{P}(\frac{1}{K}\sum_{i=1}^{K}\langle \xi_k, x^* - x_k\rangle \ge t),$$

Proof of Theorem 6

- $\langle \xi_k, x^* - x_k \rangle$ is a bounded difference martingale sequence
  - $Z_k = (x_1, \cdots, x_{k+1})$
  - Since $\mathbf{E}[\xi_k | Z_{k-1}] = 0$ and $\mathbf{E}[x_k | Z_{k-1}] = x_k$.

$$\mathbf{E} \langle \xi_k, x^* - x_k \rangle = 0.$$

  - Since $\|\xi_k\|_2 = ||g_k - f'(x_k)|| \leq 2M$

$$| \langle \xi_k, x^* - x_k \rangle | \leq \|\xi_k\|_2 ||x^* - x_k||_2 \leq 2MR$$

- By Azuma-Hoeffding Inequality,

$$\mathrm{P}(\sum_{i=1}^{K} \langle \xi_k, x^* - x_k \rangle \geq t) \leq \exp(-\frac{t^2}{2KM^2R^2}).$$

- Substituting $t = MR\sqrt{K}\epsilon$

$$\mathrm{P}(\frac{1}{K} \sum_{i=1}^{K} \langle \xi_k, x^* - x_k \rangle \geq \frac{MR\epsilon}{\sqrt{K}}) \leq \exp(-\frac{\epsilon^2}{2}).$$

Adaptive stepsizes

- choose an appropriate metric and associated distance-generating function h.
- it may be advantageous to adapt the metric being used, or at least the stepsizes, to achieve faster convergence guarantees.
- a simple scheme
$$h(x) = \frac{1}{2}x^T Ax$$
where A may change depending on information observed during solution of the problem.

Adaptive stepsizes

- Recall the bounds

$$\mathbf{E}[f(\bar{x}_K) - f(x^*)] \le \mathbf{E}[\frac{R^2}{K\alpha_k} + \frac{1}{2K}\sum_{k=1}^{K}\alpha_k\|g_k\|^2]. \qquad (34)$$

- Taking $\alpha_k = R/\sqrt{\sum_{i=1}^{k}\|g\|^2}$,

$$\mathbf{E}[f(\bar{x}_k) - f(x^*)] \le 2\frac{R}{K}\mathbf{E}[(\sum_{k=1}^{K}\|g_k\|^2)^{\frac{1}{2}}]. \qquad (35)$$

- if $\mathbf{E}[\|g_k\|^2] \le M^2$ for all k, then

$$\mathbf{E}[(\sum_{k=1}^{K}\|g_k\|^2)^{\frac{1}{2}}] \le [(\mathbf{E}\sum_{k=1}^{K}\|g_k\|^2)]^{\frac{1}{2}} \le \sqrt{M^2K} = M\sqrt{K} \qquad (36)$$

Variable metric methods

- **Variable metric methods**

$$x_{k+1} = \arg\min_{x \in C} \{ \langle g_k, x \rangle + \frac{1}{2} \langle x - x_k, H_k(x - x_k) \rangle \}$$

- **Projected subgradient method:** $H_k = \alpha_k I$,
- **Newton method:** $H_k = \nabla^2 f(x_k)$,
- **AdaGrad:** $H_k = \frac{1}{\alpha} \mathrm{diag}(\sum_{i=1}^{k} g_i . * g_i)^{\frac{1}{2}}$

Theorem 9: Convergence of Variable metric methods

Let $H_k > 0$ be a sequence of positive define matrices, where H_k is a function of $g_1, \cdots, g_k$. Let g_k be stochastic subgradient with $\mathbf{E}[g_k|x_k] \in \partial f(x_k)$. Then

$$\mathbf{E}[\sum_{k=1}^{K}(f(x_k) - f(x^*))] \leq \frac{1}{2}\mathbf{E}[\sum_{k=2}^{K}(\|x_k - x^*\|_{H_k}^2 - \|x_k - x^*\|_{H_{k-1}}^2)] + \frac{1}{2}\mathbf{E}[\|x_1 - x^*\|_{H_1}^2 + \sum_{k=1}^{K}\|g_k\|_{H_k^{-1}}^2].$$

$$\mathbf{E}[f(x_k) - f(x^*)] \leq \frac{1}{2}\mathbf{E}[\|x_k - x^*\|_{H_k}^2 - \|x_{k+1} - x^*\|_{H_k}^2 + \|g_k\|_{H_k^{-1}}^2]$$

Proof of Theorem 9

- By non-expansiveness of $\pi_C(x)$ under $\|x\|_{H_k}^2 = \langle x, H_k x\rangle$

$$\|x_{k+1} - x^*\|_{H_k}^2 \le \|x_k - H_k^{-1} g_k - x^*\|_{H_k}^2$$

- Define $\xi_k = g_k - f'(x_k)$

$$\begin{aligned} & \frac{1}{2}||x_{k+1} - x^*||_{H_k}^2 \\ \le\; & \frac{1}{2}||x_k - x^*||_{H_k}^2 + \langle g_k, x^* - x_k\rangle + \frac{1}{2}||g_k||_{H_k^{-1}}^2 \\ =\; & \frac{1}{2}||x_k - x^*||_{H_k}^2 + \langle f'(x_k), x^* - x_k\rangle + \frac{1}{2}||g_k||_{H_k^{-1}}^2 + \langle \xi_k, x^* - x_k\rangle \\ \le\; & \frac{1}{2}||x_k - x^*||_{H_k}^2 - (f(x_k) - f(x^*)) + \frac{1}{2}||g_k||_{H_k^{-1}}^2 + \langle \xi_k, x^* - x_k\rangle \end{aligned}$$

-

$$\mathbf{E}[f(x_k) - f(x^*)] \le \frac{1}{2}\mathbf{E}[\|x_k - x^*\|_{H_k}^2 - \|x_{k+1} - x^*\|_{H_k}^2 + \|g_k\|_{H_k^{-1}}^2]$$

Variable metric methods

Assume $H_k = H$ for all k. Then

$$\mathbf{E}\left[\sum_{k=1}^{K}(f(x_k) - f(x^*))\right] \le \frac{1}{2}\mathbf{E}\left[\|x_1 - x^*\|_{H_1}^2 + \sum_{k=1}^{K}\|g_k\|_{H^{-1}}^2\right]$$

Minimize the error by considering:

$$\begin{aligned} \min \quad & \sum_{t=1}^{K}\|g_t\|_{H^{-1}}^2 \\ \text{s.t.} \quad & H \succeq 0 \\ & \mathrm{tr}(H) \le c \end{aligned} \qquad H = \begin{pmatrix} s_1 & 0 & \cdots & 0 \\ 0 & s_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & s_d \end{pmatrix}$$

It is equivalent to

$$\min_s \sum_{i=1}^{d}\frac{\sum_{t=1}^{K} g_{t,i}^2}{s_i}, \qquad \text{s.t.} \quad 1^\top s \le c, \quad s \ge 0.$$

Variable metric methods

The Lagrangian function of the problem is:

$$L(s, \lambda, \theta) = \sum_{i=1}^{d} \frac{\|g_{1:K,i}\|_2^2}{s_i} - \lambda^\top s + \theta(1^\top s - c).$$

The complementarity condition gives $\lambda_i s_i = 0$. Then we obtain

$$\frac{\partial L}{\partial s_i} = -\frac{\|g_{1:K,i}\|_2^2}{s_i^2} - \lambda_i + \theta = 0,$$

which yields: $0 = -\|g_{1:K,i}\|_2^2 - \lambda_i s_i^2 + \theta s_i^2 = -\|g_{1:K,i}\|_2^2 + \theta s_i^2$. Hence, we have

$$\boxed{s_i = \frac{c\|g_{1:K,i}\|_2}{\sum_{i=1}^d \|g_{1:K,i}\|_2}}$$

Taking $c = \sum_{i=1}^d \|g_{1:K,i}\|_2$ gives $\boxed{s_i = \|g_{1:K,i}\|_2}$.

Corollary: Convergence of AdaGrad

Let $R_\infty = \sup_{x\in C} \|x - x^*\|_\infty$ and let the conditions of Theorem 9 hold. Then we have

$$\mathbf{E}[\sum_{k=1}^{K}(f(x_k) - f(x^*))] \le \frac{1}{2\alpha} R_\infty^2 \mathbf{E}[\mathrm{tr}(M_K)] + \alpha \mathbf{E}[\mathrm{tr}(M_K)]$$

where $M_k = \mathrm{diag}(\sum_{i=1}^{k} g_i .* g_i)^{\frac{1}{2}}$ and $H_k = \frac{1}{\alpha} M_k$

- Let $\alpha = \mathbb{R}_\infty$, Then

$$\mathbf{E}[f(\bar{x}_K) - f(x^*)] \le \frac{3}{2K} R_\infty \mathbf{E}[\mathrm{tr}(M_K)] = \frac{3}{2K} R_\infty \sum_{j=1}^{n} \mathbf{E}[(\sum_{k=1}^{K} g_{k,j}^2)^{\frac{1}{2}}] \tag{37}$$

- If $C = \{x : \|x\| \le 1\}$, the bound is lower than the one of adaptive stepsize.

Proof of Corolory

- Aim: $\|x_k - x^*\|_{H_k}^2 - \|x_k - x^*\|_{H_{k-1}}^2 \leq ||x_k - x^*||_\infty^2 tr(H_k - H_{k-1})$
  Let $z = x - x^*$

$$\begin{aligned}
& \|z\|_{H_k}^2 - \|z\|_{H_{k-1}}^2 \\
= & \sum_{j=1}^{n} H_{k,j} z_j^2 - \sum_{j=1}^{n} H_{k-1,j} z_j^2 \\
= & \sum_{j=1}^{n} (H_{k,j} - H_{k-1,j}) z_j^2 \\
\leq & ||z||_\infty^2 \sum_{j=1}^{n} (H_{k,j} - H_{k-1,j}) \\
= & ||z||_\infty^2 tr(H_k - H_{k-1})
\end{aligned}$$

Proof of Corolory

- Assume $a = (a_1, a_2, ..., a_T)$, a simple inequality(prove by induction),

$$\sum_{t=1}^{T} \frac{a_t^2}{\sqrt{a_1^2 + \cdots + a_t^2}} \le 2\sqrt{a_1^2 + \cdots + a_T^2}$$

- Aim: $\sum_{k=1}^{K} \|g_k\|_{H_k^{-1}} \le 2\alpha \mathrm{tr}(M_K)$.

$$\begin{aligned} & \sum_{k=1}^{K} \|g_k\|_{H_k^{-1}}^2 \\ = \; & \alpha \sum_{k=1}^{K} \sum_{j=1}^{n} \frac{g_{k,j}^2}{M_{k,j}} = \alpha \sum_{j=1}^{n} \sum_{k=1}^{K} \frac{g_{k,j}^2}{\sqrt{\sum_{i=1}^{k} g_{i,j}^2}} \\ \le \; & 2\alpha \sum_{j=1}^{n} \sqrt{\sum_{i=1}^{K} g_{i,j}^2} = 2\alpha \mathrm{tr}(M_K) \end{aligned} \tag{38}$$

Summary

- expectation

$$\mathbf{E}[f(\bar{x}_K) - f(x))] \leq \frac{3RM}{2\sqrt{K}}$$

- convergence in probability

$$f(\bar{x}_K) - f(x^*) \leq \frac{3RM}{2\sqrt{K}} + \frac{RM\sqrt{2\log\frac{1}{\delta}}}{\sqrt{K}}.$$

  with probability at least $1 - \delta$

- Using proper metric and adapted strategy can improve the convergence: Mirror Descent method and Adagrad.

Outline

Gradient methods

- Rewrite the ERM problem

$$\min_{x\in\mathbb{R}^n} f(x) := \frac{1}{n}\sum_{i=1}^{n} f_i(x) \tag{39}$$

- **gradient methods**

$$x_{k+1} = x_k - \alpha_k \nabla f(x_k) \tag{40}$$

- the update is equal to

$$x_{k+1} = \arg\min_{x} \quad f(x_k) + \langle \nabla f(x_k), x - x_k\rangle + \frac{1}{2\alpha_k}||x - x_k||_2^2 \tag{41}$$

Basic Properties

- We only consider the convex defferentiable functions.
- **convex functions**:

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda)f(y), \forall \lambda \in [0,1], x, y$$

- M**-Lipschitz functions**:

$$|f(x) - f(y)| \leq M||x-y||_2$$

- L**-smooth functions**:

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x-y\|_2$$

- μ**-strongly convex functions**:

$$f(\lambda x+(1-\lambda)y) \leq \lambda f(x)+(1-\lambda)f(y)-\frac{\mu}{2}\lambda(1-\lambda)||x-y||_2^2, \forall \lambda \in [0,1], x, y$$

Some useful results

- **convex functions**:

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

- M**-Lipschitz functions**:

$$\|\nabla f(x)\|_2 \leq M$$

- L**-smooth functions**: $\frac{L}{2}x^T x - f(x)$ is convex

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|x - y\|_2^2$$

$$\frac{1}{2L}\|\nabla f(x)\|_2^2 \leq f(x) - f(x^*) \leq \frac{L}{2}\|x - x^*\|_2^2$$

- μ**-strongly convex functions**: $f(x) - \frac{\mu}{2}x^T x$ is convex

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2}\|x - y\|_2^2$$

Co-coercivity of gradient

if f is convex with $\mathbf{dom}\, f = \mathbf{R}^n$ and $(L/2)x^\top x - f(x)$ is convex then

$$(\nabla f(x) - \nabla f(y))^\top (x - y) \geq \frac{1}{L} \|\nabla f(x) - \nabla f(y)\|_2^2 \quad \forall x, y$$

proof: define convex functions f_x, f_y with domain $\mathbf{R}^n$:

$$f_x(z) = f(z) - \nabla f(x)^\top z, \quad f_y(z) = f(z) - \nabla f(y)^\top z$$

the functions $(L/2)z^\top z - f_x(z)$ and $(L/2)z^\top z - f_y(z)$ are convex

- $z = x$ minimizes $f_x(z)$; from the left-hand inequality,

$$\begin{aligned} f(y) - f(x) - \nabla f(x)^\top (y - x) &= f_x(y) - f_x(x) \\ &\geq \frac{1}{2L} \|\nabla f_x(y)\|_2^2 = \frac{1}{2L} \|\nabla f(y) - \nabla f(x)\|_2^2 \end{aligned}$$

- similarly, $z = y$ minimizes $f_y(z)$; therefore

$$f(x) - f(y) - \nabla f(y)^\top (x - y) \geq \frac{1}{2L} \|\nabla f(y) - \nabla f(x)\|_2^2$$

combining the two inequalities shows co-coercivity

Extension of co-coercivity

if f is strongly convex and ∇f is Lipschitz continuous, then

$$g(x) = f(x) - \frac{\mu}{2}\|x\|_2^2$$

is convex and ∇g is Lipschitz continuous with parameter $L - \mu$.

co-coercivity of g gives

$$\begin{aligned}&(\nabla f(x) - \nabla f(y))^\top (x - y)\\ &\geq \frac{\mu L}{\mu + L}\|x - y\|_2^2 + \frac{1}{\mu + L}\|\nabla f(x) - \nabla f(y)\|_2^2\end{aligned}$$

for all $x, y \in \textbf{dom}\, f$

Convergence guarantees

Assumption

- f is L-smooth and μ-strongly convex.

lemma: Coercivity of gradients

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \frac{L\mu}{L+\mu}\|x-y\|^2 + \frac{1}{L+\mu}\|\nabla f(x) - \nabla f(y)\|^2 \quad (42)$$

Theorem: Convergence rates of GD

Let $\alpha_k = \frac{2}{L+\mu}$ and let $\kappa = \frac{L}{\mu}$. Define $\Delta_k = \|x_k - x^*\|$. Then we get,

$$f(x_{T+1}) - f(x^*) \leq \frac{L\Delta_1^2}{2}\exp(-\frac{4T}{\kappa+1}). \quad (43)$$

Proof of Theorem

- $$\begin{aligned}\Delta_{k+1}^2 &= ||x_{k+1} - x^*||_2^2 = ||x_k - \alpha_k \nabla f(x_k) - x^*||_2^2 \\ &= ||x_k - x^*||_2^2 - 2\alpha_k \langle \nabla f(x_k), x_k - x^* \rangle + \alpha_k^2 ||\nabla f(x_k)||_2^2 \\ &= \Delta_k^2 - 2\alpha_k \boxed{\langle \nabla f(x_k), x_k - x^* \rangle} + \alpha_k^2 ||\nabla f(x_k)||_2^2\end{aligned}$$

- By the lemma

$$\begin{aligned}\Delta_{k+1}^2 &\leq \Delta_k^2 - 2\alpha_k \boxed{\left(\frac{L\mu}{L+\mu}\Delta_k^2 + \frac{1}{L+\mu}\|\nabla f(x)\|^2\right)} + \alpha_k^2 ||\nabla f(x_k)||_2^2 \\ &= (1 - 2\alpha_k \frac{L\mu}{L+\mu})\Delta_k^2 + (-\frac{2\alpha_k}{L+\mu} + \alpha_k^2)||\nabla f(x_k)||_2^2 \\ &\leq (1 - 2\alpha_k \frac{L\mu}{L+\mu})\Delta_k^2 + (-\frac{2\alpha_k}{L+\mu} + \alpha_k^2)L^2\Delta_k^2 \qquad (44)\end{aligned}$$

Proof of Theorem

- $\alpha_k = \frac{2}{L+\mu}$

$$\begin{aligned} \Delta_{k+1}^2 &\le (1 - \frac{4L\mu}{(L+\mu)^2})\Delta_k^2 \\ &= (\frac{L-\mu}{L+\mu})^2\Delta_k^2 = (\frac{\kappa-1}{\kappa+1})^2\Delta_k^2 \end{aligned}$$

-

$$\begin{aligned} \Delta_{T+1}^2 &\le (\frac{\kappa-1}{\kappa+1})^{2T}\Delta_1^2 \\ &= \Delta_1^2 \exp(2T\log(1-\frac{2}{\kappa+1})) \\ &\le \Delta_1^2 \exp(-\frac{4T}{\kappa+1}) \end{aligned}$$

-

$$f(x_{T+1}) - f(x^*) \le \frac{L}{2}\Delta_{T+1}^2 \le \frac{L\Delta_1^2}{2}\exp(-\frac{4T}{\kappa+1})$$

Stochastic Gradient methods

- ERM problem

$$\min_{x\in\mathbb{R}^n} f(x) := \frac{1}{n}\sum_{i=1}^{n} f_i(x) \tag{45}$$

- **gradient descent**

$$x_{k+1} = x_k - \alpha_k \nabla f(x_k) \tag{46}$$

- **stochastic gradient descent**

$$x_{k+1} = x_k - \alpha_k \nabla f_{s_k}(x_k), \tag{47}$$

where s_k is uniformly sampled from $\{1, \cdots, n\}$

Convergence guarantees

Assumption

- $f(x)$ is L-smooth: $||\nabla f(x) - \nabla f(y)||_2^2 \leq L||x-y||_2^2$
- $f(x)$ is μ-strongly convex: $\langle \nabla f(x) - \nabla f(y), x-y \rangle \geq \mu||x-y||_2^2$
- $\mathbf{E}_s[\nabla f_s(x)] = \nabla f(x)$
- $\mathbf{E}_s||\nabla f_s(x)||^2 \leq M^2$

Theorem: Convergence rates of SGD

Define $\Delta_k = \|x_k - x^*\|$. For a fixed Stepsize $\alpha_k = \alpha$, $0 < \alpha < \frac{1}{2\mu}$ we have,

$$\mathbf{E}[f(x_{T+1}) - f(x^*)] \leq \frac{L}{2}\mathbf{E}[\Delta_{T+1}^2] \leq \frac{L}{2}[(1-2\alpha\mu)^T \Delta_1^2 + \frac{\alpha M^2}{2\mu}]. \tag{48}$$

- $$\begin{aligned} \Delta_{k+1}^2 &= ||x_{k+1} - x^*||_2^2 = ||x_k - \alpha_k \nabla f_{s_k}(x_k) - x^*||_2^2 \\ &= ||x_k - x^*||_2^2 - 2\alpha_k \langle \nabla f_{s_k}(x_k), x_k - x^* \rangle + \alpha_k^2 ||\nabla f_{s_k}(x_k)||_2^2 \\ &= \Delta_k^2 - 2\alpha_k \langle \nabla f_{s_k}(x_k), x_k - x^* \rangle + \alpha_k^2 ||\nabla f_{s_k}(x_k)||_2^2 \end{aligned}$$

- Using $E[X] = E[E[X|Y]]$:

$$\begin{aligned} & \mathbf{E}_{s_1,\ldots,s_k}[\langle \nabla f_{s_k}(x_k), x_k - x^* \rangle] = \mathbf{E}_{s_1,\ldots,s_{k-1}}[\mathbf{E}_{s_k}[\langle \nabla f_{s_k}(x_k), x_k - x^* \rangle]] \\ = \; & \mathbf{E}_{s_1,\ldots,s_{k-1}}[\langle \mathbf{E}_{s_k}[\nabla f_{s_k}(x_k)], x_k - x^* \rangle] \\ = \; & \mathbf{E}_{s_1,\ldots,s_{k-1}}[\langle \nabla f(x_k), x_k - x^* \rangle] \\ = \; & \mathbf{E}_{s_1,\ldots,s_k}[\langle \nabla f(x_k), x_k - x^* \rangle] \end{aligned}$$

- By the strongly convexity

$$\mathbf{E}_{s_1,\ldots,s_k}(\Delta_{k+1}^2) \leq (1 - 2\alpha\mu)\mathbf{E}_{s_1,\ldots,s_k}(\Delta_k^2) + \alpha^2 M^2 \tag{49}$$

Proof of Theorem

- Taking induction from $k = 1$ to $k = T$, we have

$$\mathbf{E}_{s_1,\ldots,s_T}(\Delta_{T+1}^2) \leq (1-2\alpha\mu)^T \Delta_1^2 + \sum_{i=0}^{T-1} (1-2\alpha\mu)^i \alpha^2 M^2 \quad (50)$$

- under the assumption that $0 \leq 2\alpha\mu \leq 1$, we have

$$\sum_{i=0}^{\infty} (1-2\alpha\mu)^i = \frac{1}{2\alpha\mu}$$

- Then

$$\mathbf{E}_{s_1,\ldots,s_T}(\Delta_{T+1}^2) \leq (1-2\alpha\mu)^T \Delta_1^2 + \frac{\alpha M^2}{2\mu} \quad (51)$$

Convergence guarantees

- For fixed stepsize, we don't have the convergence
- For diminishing stepsize, the order of convergence is $O(\frac{1}{T})$

Theorem: Convergence rates of SGD

Define $\Delta_k = \|x_k - x^*\|$. For a diminishing stepsize

$$\alpha_k = \frac{\beta}{k+\gamma} \text{ for some } \beta > \frac{1}{2\mu} \text{ and } \gamma > 0 \text{ such that } \alpha_1 \leq \frac{1}{2\mu}.$$

Then we have, for any $T \geq 1$

$$\mathbf{E}[f(x_T) - f(x^*)] \leq \frac{L}{2}\mathbf{E}[\Delta_T^2] \leq \frac{L}{2}\frac{v}{\gamma + T}, \tag{52}$$

where $v = \max(\frac{\beta^2 M^2}{2\beta\mu - 1}, (\gamma + 1)\Delta_1^2)$

Proof of Theorem

- Recall the bounds

$$\mathbf{E}_{s_1,\ldots,s_k}(\Delta_{k+1}^2) \quad \leq \quad (1-2\alpha\mu)\mathbf{E}_{s_1,\ldots,s_k}(\Delta_k^2) + \alpha^2 M^2 \qquad (53)$$

- We prove it by induction. Firstly, the definition of v ensures that it holds for k = 1.
- Assume the conclusion holds for some k, it follows that

$$\begin{aligned}
\mathbf{E}(\Delta_{k+1}^2) &\leq (1-\frac{2\beta\mu}{\hat{k}})\frac{v}{\hat{k}} + \frac{\beta^2 M^2}{\hat{k}^2} \quad (\text{ with } \hat{k} := \gamma + k) \\
&= (\frac{\hat{k}-2\beta\mu}{\hat{k}^2})v + \frac{\beta^2 M^2}{\hat{k}^2} \\
&= \frac{\hat{k}-1}{\hat{k}^2}v \boxed{-\frac{2\beta\mu - 1}{\hat{k}^2}v + \frac{\beta^2 M^2}{\hat{k}^2}} \\
&\leq \frac{v}{\hat{k}+1}
\end{aligned}$$

stochastic optimization

- stochastic subgradient descent: $O(1/\epsilon^2)$
- stochastic gradient descent with strong convexity $O(1/\epsilon)$
- stochastic gradient descent with strong convexity and smoothness $O(1/\epsilon)$

deterministic optimization

- subgradient descent: $O(n/\epsilon^2)$
- gradient descent with strong convexity $O(n/\epsilon)$
- gradient descent with strong convexity and smoothness $O(n\log(1/\epsilon))$

The complexity refers to the times of computation of component (sub)gradients. We need to compute n gradients in every iterations of GD and one gradient in SGD.

Outline

Variance Reduction

Assumption

- $f(x)$ is L-smooth: $||\nabla f(x) - \nabla f(y)||_2 \leq L||x - y||_2$
- $f(x)$ is μ-strongly convex: $\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \mu ||x - y||_2^2$
- $\mathbf{E}_s[\nabla f_s(x)] = \nabla f(x)$
- $\mathbf{E}_s||\nabla f_s(x)||^2 \leq M^2$

- GD: linear convergence $O(n \log(1/\epsilon))$
- SGD: sublinear convergence $O(1/\epsilon)$

What is the essential difference between SGD and GD?

Variance Reduction

- **GD**

$$\begin{aligned}\Delta_{k+1}^2 &= ||x_{k+1} - x^*||_2^2 = ||x_k - \alpha\nabla f(x_k) - x^*||_2^2 \\ &= \Delta_k^2 - 2\alpha\langle\nabla f(x_k), x_k - x^*\rangle + \alpha^2||\nabla f(x_k)||_2^2 \\ &\leq (1-2\alpha\mu)\Delta_k^2 + \alpha^2||\nabla f(x_k)||_2^2 \quad (\mu - \text{strongly convex}) \\ &\leq (1-2\alpha\mu+\alpha^2L^2)\Delta_k^2 \quad (L - \text{smooth})\end{aligned}$$

- **SGD**

$$\begin{aligned}\mathbf{E}\Delta_{k+1}^2 &= \mathbf{E}||x_{k+1} - x^*||_2^2 = \mathbf{E}||x_k - \alpha\nabla f_{s_k}(x_k) - x^*||_2^2 \\ &= \mathbf{E}\Delta_k^2 - 2\alpha\mathbf{E}\langle\nabla f_{s_k}(x_k), x_k - x^*\rangle + \alpha^2\mathbf{E}||\nabla f_{s_k}(x_k)||_2^2 \\ &= \mathbf{E}\Delta_k^2 - 2\alpha\mathbf{E}\langle\nabla f(x_k), x_k - x^*\rangle + \alpha^2\mathbf{E}||\nabla f_{s_k}(x_k)||_2^2 \\ &\leq (1-2\alpha\mu)\mathbf{E}\Delta_k^2 + \alpha^2\mathbf{E}||\nabla f_{s_k}(x_k)||_2^2 \quad (\mu - \text{strongly convex}) \\ &\leq (1-2\alpha\mu)\mathbf{E}\Delta_k^2 + \alpha^2\mathbf{E}||\nabla f_{s_k}(x_k) - \nabla f(x_k) + \nabla f(x_k)||_2^2 \\ &\leq (1-2\alpha\mu+2\alpha^2L^2)\mathbf{E}\Delta_k^2 + \boxed{2\alpha^2\mathbf{E}||\nabla f_{s_k}(x_k) - \nabla f(x_k)||_2^2}\end{aligned}$$

Variance Reduction

$$\mathbf{E}\Delta_{k+1}^2 \leq \underbrace{(1 - 2\alpha\mu + 2\alpha^2 L^2)\mathbf{E}\Delta_k^2}_{A} + \underbrace{\boxed{2\alpha^2 \mathbf{E}||\nabla f_{s_k}(x_k) - \nabla f(x_k)||_2^2}}_{B} \quad (54)$$

- a worst case convergence rate of $\sim 1/T$ for SGD
- In practice, the actual convergence rate may be somewhat better than this bound.
- Initially, $B << A$ and we observe the linear rate regime, once $B > A$ we observe $1/T$ rate.
- How to reduce variance term B to speed up SGD?
  - SAG (Stochastic average gradient)
  - SAGA
  - SVRG (Stochastic variance reduced gradient)

SAG method

- SAG method (Le Roux, Schmidt, Bach 2012)

$$x_{k+1} = x_k - \frac{\alpha_k}{n}\sum_{i=1}^{n} g_k^i = x_k - \alpha_k\left(\frac{1}{n}(\nabla f_{s_k}(x_k) - g_{k-1}^{s_k}) + \frac{1}{n}\sum_{i=1}^{n} g_{k-1}^i\right) \tag{55}$$

  where

$$g_k^i = \begin{cases} \nabla f_i(x_k) & \text{if } i = s_k, \\ \nabla g_{k-1}^i & o.w., \end{cases} \tag{56}$$

  and s_k is uniformly sampled from $\{1, \cdots, n\}$
- complexity($\#$ component gradient evaluations): $O(\max\{n, \frac{L}{\mu}\}\log(1/\epsilon))$
- need to store most recent gradient of each component.
- SAGA(Defazio, Bach,Julien 2014) is unbaised revision of SAG

$$x_{k+1} = x_k - \alpha_k(\nabla f_{i_k}(x_k) - g_{k-1}^{i_k} + \frac{1}{n}\sum_{i=1}^{n} g_{k-1}^i) \tag{57}$$

SVRG method

- SVRG method (Johnson and Zhang 2013)

$$
\begin{aligned}
v_k &= \nabla f_{s_k}(x_k) - \nabla f_{s_k}(y) + \nabla f(y) \\
x_{k+1} &= x_k - \alpha_k v_k
\end{aligned}
$$

  where and s_k is uniformly sampled from $\{1, \cdots, n\}$

- v_k is unbiased estimation of gradient $\nabla f(x_k)$

$$
\mathbf{E}v_k = \nabla f(x_k) + \nabla f(y) - \nabla f(y) = \nabla f(x_k). \tag{58}
$$

- Recall the bound

$$
\mathbf{E}\Delta_{k+1}^2 \leq (1 - 2\alpha\mu)\mathbf{E}\Delta_k^2 + \alpha^2 \mathbf{E}||v_k||_2^2 \tag{59}
$$

SVRG method

- Additional assumption: $L - smoothness$ for component functions

$$\|\nabla f_i(x) - \nabla f_i(y)\|_2 \leq L\|x - y\|_2 \tag{60}$$

- Let's analyze the "variance"

$$\begin{aligned}
& \mathbf{E}\|v_k\|_2^2 \\
= \;& \mathbf{E}\|\nabla f_{s_k}(x_k) - \nabla f_{s_k}(y) + \nabla f(y)\|_2^2 \\
= \;& \mathbf{E}\|\nabla f_{s_k}(x_k) - \nabla f_{s_k}(y) + \nabla f(y) + \nabla f_{s_k}(x^*) - \nabla f_{s_k}(x^*)\|_2^2 \\
\leq \;& 2\mathbf{E}\|\nabla f_{s_k}(x_k) - \nabla f_{s_k}(x^*)\|_2^2 + 2\mathbf{E}\|\nabla f_{s_k}(y) - \nabla f(y) - \nabla f_{s_k}(x^*)\|_2^2 \\
\leq \;& 2L^2\mathbf{E}\Delta_k^2 + 2\mathbf{E}\|\nabla f_{s_k}(y) - \nabla f_{s_k}(x^*)\|_2^2 \\
\leq \;& 2L^2\mathbf{E}\Delta_k^2 + 2L^2\mathbf{E}\|y - x^*\|^2
\end{aligned}$$

- if x_k and y is close to x^*, the variance is small.

SVRG method

- We only need to choose a current point as y.
- picking a fresh y more often should decrease the variance, however doing this too often involves computing too many full gradients
- Let's set $y = x_1$,

$$\mathbf{E}\Delta_{k+1}^2 \leq (1 - 2\alpha\mu + 2\alpha^2 L^2)\mathbf{E}\Delta_k^2 + 2\alpha^2 L^2 \mathbf{E}\Delta_1^2 \tag{61}$$

- Unrolling this:

$$\begin{aligned} &\mathbf{E}\Delta_{k+1}^2 \\ \leq\ & (1 - 2\alpha\mu + 2\alpha^2 L^2)\mathbf{E}\Delta_1^2 + \sum_{i=0}^{k-1}(1 - 2\alpha\mu + 2\alpha^2 L^2)^i 2\alpha^2 L^2 \mathbf{E}\Delta_1^2 \\ \leq\ & (1 - 2\alpha\mu + 2\alpha^2 L^2)^k \mathbf{E}\Delta_1^2 + 2k\alpha^2 L^2 \mathbf{E}\Delta_1^2 \end{aligned} \tag{62}$$

SVRG method

- Unrolling this:

$$\mathbf{E}\Delta_{k+1}^2 \leq (1 - 2\alpha\mu + 2\alpha^2 L^2)^k \mathbf{E}\Delta_1^2 + 2k\alpha^2 L^2 \mathbf{E}\Delta_1^2 \tag{63}$$

- Suppose we would like this to be $\leq 0.5E\Delta_1$ after T iterations.
- We pick $\alpha = O(1)\frac{\mu}{L^2}$, then it turns out that we can set $T = O(1)\frac{L^2}{\mu^2}$.
- In fact, we can improve it to $T = O(1)\frac{L}{\mu}$.
- condition number $\kappa = \frac{L}{\mu}$

SVRG method

Algorithm 2: SVRG method

Input: $\tilde{x}_0$, α, m

for $e = 1 : E$ **do**

- $y \leftarrow \tilde{x}_{e-1}$, $x_1 \leftarrow \tilde{x}_{e-1}$.
- $g \leftarrow \nabla f(y)$ (full gradient)
- **for** $k = 1 : m$ **do**
  - pick $s_k \in \{1, \cdots, n\}$ uniformly at random.
  - $v_k = \nabla f_{s_k}(x_k) - \nabla f_{s_k}(y) + \nabla f(y)$
  - $x_{k+1} = x_k - \alpha v_k$
- **end for**
- $\boxed{\tilde{x}_e \leftarrow \frac{1}{m}\sum_{k=1}^{m} x_k}$

end for

SVRG method

Convergence of SVRG method

Suppose $0 < \alpha \leq \frac{1}{2L}$ and m sufficiently large such that

$$\rho = \frac{1}{\mu\alpha(1-2L\alpha)m} + \frac{2L\alpha}{1-2L\alpha} < 1 \tag{64}$$

then we have linear convergence in expectation

$$Ef(\tilde{x}_s) - f(x^*) \leq \rho^s[f(\tilde{x}_0) - f(x^*)] \tag{65}$$

- if $\alpha = \frac{\theta}{L}$, then

$$\rho = \frac{L/\mu}{\theta(1-2\theta)m} + \frac{2\theta}{1-2\theta} \tag{66}$$

  choosing $\theta = 0.1$ and $m = 50(L/\mu)$ results in $\rho = 0.5$
- overall complexity: $O((\frac{L}{\mu} + n)\log(1/\epsilon))$

- $$\begin{aligned}\mathbf{E}\Delta_{k+1}^2 &= \mathbf{E}||x_{k+1}-x^*||_2^2 = \mathbf{E}||x_k-\alpha v_k - x^*||_2^2 \\ &= \mathbf{E}\Delta_k^2 - 2\alpha\mathbf{E}\left\langle v_k, x_k - x^*\right\rangle + \alpha^2\mathbf{E}||v_k||_2^2 \\ &= \mathbf{E}\Delta_k^2 - 2\alpha\mathbf{E}\left\langle \nabla f(x_k), x_k - x^*\right\rangle + \alpha^2\mathbf{E}||v_k||_2^2 \\ &\leq \mathbf{E}\Delta_k^2 - 2\alpha\mathbf{E}(f(x_{k-1}) - f(x^*)) + \alpha^2\boxed{E||v_k||_2^2}\end{aligned}$$

- By smoothness of $f_i(x)$

$$\|\nabla f_i(x) - \nabla f_i(x^*)\|^2 \leq 2L[f_i(x) - f_i(x^*) - \nabla f_i(x^*)^T(x-x^*)] \quad (67)$$

- summing above inequalities over $1,2,\cdots,n$ and using $\nabla f(x^*) = 0$

$$\frac{1}{n}\sum_{i=1}^{n}\|\nabla f_i(x) - \nabla f_i(x^*)\|^2 \leq 2L[f(x) - f(x^*)] \quad (68)$$

proof of theorem

- Using $E[(X-E[X])^2] \leq E[X^2]$, we obtain the bound

$$\begin{aligned}
& \mathbf{E}_s||v_k||_2^2 \\
= \; & \mathbf{E}\|\nabla f_{s_k}(x_k) - \nabla f_{s_k}(y) + \nabla f(y) + \nabla f_{s_k}(x^*) - \nabla f_{s_k}(x^*)\|_2^2 \\
\leq \; & 2\mathbf{E}\|\nabla f_{s_k}(x_k) - \nabla f_{s_k}(x^*)\|_2^2 + 2\mathbf{E}\|\nabla f_{s_k}(y) - \nabla f(y) - \nabla f_{s_k}(x^*)\|_2^2 \\
= \; & 2\mathbf{E}\|\nabla f_{s_k}(x_k) - \nabla f_{s_k}(x^*)\|_2^2 \\
& +2\mathbf{E}\|\nabla f_{s_k}(y) - \nabla f_{s_k}(x^*) - \mathbf{E}[\nabla f_{s_k}(y) - \nabla f_{s_k}(x^*)]\|_2^2 \\
\leq \; & 2\mathbf{E}\|\nabla f_{s_k}(x_k) - \nabla f_{s_k}(x^*)\|_2^2 + 2\mathbf{E}\|\nabla f_s(y) - \nabla f_s(x^*)\|^2 \\
\leq \; & 4L[f(x_k) - f(x^*) + f(y) - f(x^*)]
\end{aligned}$$

- now continue the derivation

$$\begin{aligned}
& \mathbf{E}\Delta_{k+1}^2 \leq \mathbf{E}\Delta_k^2 - 2\alpha\mathbf{E}(f(x_k) - f(x^*)) + \alpha^2 \boxed{E||v_k||_2^2} \\
\leq \; & \mathbf{E}\Delta_k^2 - 2\alpha(1-2\alpha L)\mathbf{E}(f(x_k) - f(x^*)) + 4L\alpha^2[f(y) - f(x^*)]
\end{aligned}$$

proof of theorem

- summing over $k = 1, \cdots, m$ (note that $y = \tilde{x}_{e-1}$ and $\tilde{x}_e = \frac{1}{m}\sum_{k=1}^{m} x_k$)

$$\begin{aligned} & \mathbf{E}\Delta_{k+1}^2 + 2\alpha(1-2\alpha L)\sum_{k=1}^{m}\mathbf{E}(f(x_k) - f(x^*)) \\ \leq\ & \mathbf{E}\|\tilde{x}_{e-1} - x^*\|^2 + 4L\alpha^2 m\mathbf{E}[f(\tilde{x}_{e-1}) - f(x^*)] \\ \leq\ & \frac{2}{\mu}\mathbf{E}[f(\tilde{x}_{e-1}) - f(x^*)] + 4L\alpha^2 m\mathbf{E}[f(\tilde{x}_{e-1}) - f(x^*)] \end{aligned}$$

- therefore, for each stage s

$$\begin{aligned} & \mathbf{E}[f(\tilde{x}_e) - f(x^*)] \\ \leq\ & \frac{1}{m}\sum_{k=1}^{m}\mathbf{E}(f(x_k) - f(x^*)) \\ \leq\ & \frac{1}{2\alpha(1-2\alpha L)m}\left(\frac{2}{\mu} + 4mL\alpha^2\right)\mathbf{E}[f(\tilde{x}_{e-1}) - f(x^*)] \qquad (69) \end{aligned}$$

Summary

- condition number: $\kappa = \frac{L}{\mu}$
- **SVRG**: $E \sim \log(\frac{1}{\epsilon})$ so the complexity is $O((n+\kappa)\log(\frac{1}{\epsilon}))$
- **GD**: $T \sim \kappa \log(\frac{1}{\epsilon})$ so the complexity is $O(n\kappa \log(\frac{1}{\epsilon}))$
- **SGD**: $T \sim \frac{\kappa}{\epsilon}$ so the complexity is $O(\frac{\kappa}{\epsilon})$
- even though we are allowing ourselves a few gradient computations here, we don't really pay too much in terms of complexity.

Outline

Stochastic Algorithms in Deep learning

Consider problem $\min_{x\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^n f_i(x)$
References: chapter 8 in
http://www.deeplearningbook.org/

- Gradient descent

$$x^{t+1} = x^t - \frac{\alpha^t}{n}\sum_{i=1}^n \nabla f_i(x^t)$$

- Stochastic gradient descent

$$x^{k+1} = x^t - \alpha^t \nabla f_i(x^t)$$

- SGD with momentum

$$v^{t+1} = \mu^t v^t - \alpha^t \nabla f_i(x^t)$$
$$x^{t+1} = x^t + v^{t+1}$$

Stochastic Algorithms in Deep learning

- Nesterov accelerated gradient (original version)

$$
\begin{aligned}
v^{t+1} &= (1 + \mu^t)x^t - \mu^t x^{t-1} \\
x^{t+1} &= v^{t+1} - \alpha^t \nabla f_i(v^{t+1})
\end{aligned}
$$

  here $\mu^t = \frac{t+2}{t+5}$ and α^t fixed or determined by line search (inverse of Lipschitz constant).

- Nesterov accelerated gradient (momentum version)

$$
\begin{aligned}
v^{t+1} &= \mu^t v^t - \alpha^t \nabla f_i(x^t + \mu^t v^t) \\
x^{t+1} &= x^t + v^{t+1}
\end{aligned}
$$

  here $\mu^t = \frac{t+2}{t+5}$ and α^t fixed or determined by line search.

Stochastic Algorithms in Deep learning

- Adaptive Subgradient Methods (Adagrad): let $g_t = \nabla f_i(x^t)$, $g_t^2 = \mathrm{diag}[g_t g_t^T] \in \mathbb{R}^d$, and initial $G_1 = g_1^2$. At step t

$$x^{t+1} = x^t - \frac{\alpha^t}{\sqrt{G^t + \epsilon \mathbf{1}_d}} \nabla f_i(x^t)$$
$$G^{t+1} = G^t + g_{t+1}^2$$

in the upper and the following iterations we use element-wise vector-vector multiplication.

- Adam: initial $E[g^2]_0 = 0$, $E[g]_0 = 0$. At step t,

$$
\begin{aligned}
E[g]_t &= \mu E[g]_{t-1} + (1-\mu) g_t \\
E[g^2]_t &= \rho E[g^2]_{t-1} + (1-\rho) g_t^2 \\
\widehat{E}[g]_t &= \frac{E[g]_t}{1-\mu^t} \\
\widehat{E}[g^2]_t &= \frac{E[g^2]_t}{1-\rho^t} \\
x^{t+1} &= x^t - \frac{\alpha}{\sqrt{\widehat{E}[g^2]_t + \epsilon \mathbf{1}_d}} \widehat{E}[g]_t
\end{aligned}
$$

here ρ, μ are decay rates, α is learning rate.

Optimization algorithms in Deep learning

随机梯度类算法

- pytorch/caffe2 里实现的算法有adadelta, adagrad, adam, nesterov, rmsprop, YellowFin
  https://github.com/pytorch/pytorch/tree/master/caffe2/sgd
- pytorch/torch 里有： sgd, asgd, adagrad, rmsprop, adadelta, adam, adamax
  https://github.com/pytorch/pytorch/tree/master/torch/optim
- tensorflow 实现的算法有： Adadelta, AdagradDA, Adagrad, ProximalAdagrad, Ftrl, Momentum, adam, Momentum, CenteredRMSProp
  具体实现:
  https://github.com/tensorflow/tensorflow/blob/master/tensorflow/core/kernels/training_ops.cc

数值例子：逻辑回归

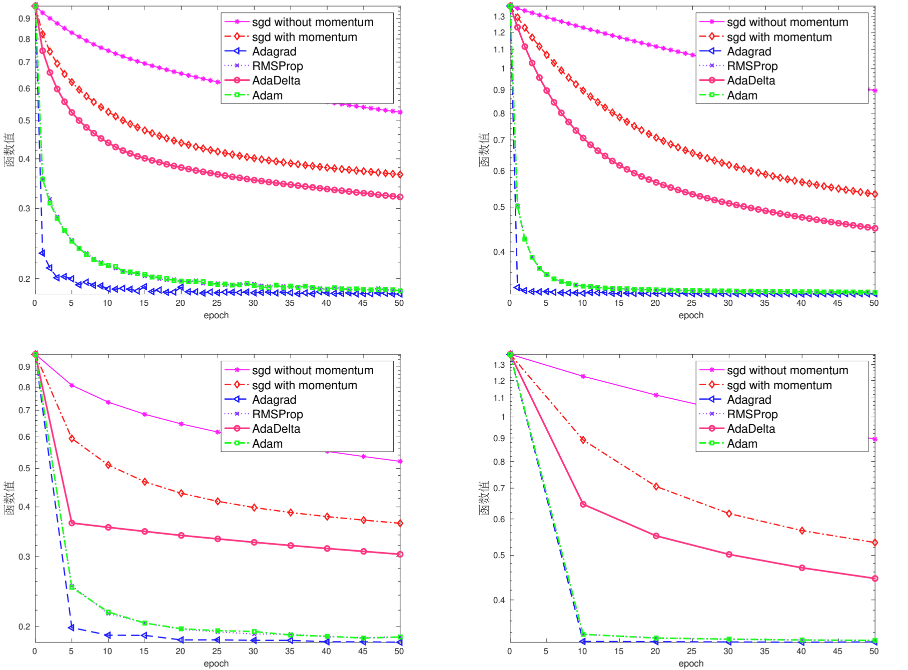
给定数据集$\{(a_i, b_i)_{i=1}^N\}$，逻辑回归对应的优化问题可以写成如下形式

$$\min_{x\in\mathbb{R}^n} \quad f(x) = \frac{1}{N}\sum_{i=1}^N f_i(x) = \frac{1}{N}\sum_{i=1}^N \log(1+\exp(-b_i\cdot a_i^\top x)) + \lambda\|x\|_2^2,$$

每步我们随机取一个数据i_k对应的梯度$\nabla f_{i_k}(x^k)$作随机梯度下降，其迭代格式可以写成

$$x^{k+1} = x^k - \tau_k \nabla f_{i_k}(x^k) = x^k - \tau_k \frac{1}{N}\left(\frac{-\exp(-b_{i_k}\cdot a_{i_k}^\top x^k)b_{i_k}a_{i_k}}{1+\exp(-b_{i_k}\cdot a_{i_k}^\top x^k)} + 2\lambda x^k\right),$$

其中i_k为从$\{1,\cdots,N\}$随机抽取的一个样本，τ_k 为步长。采用LIBSVM网站的数据集，并令$\lambda = 10^{-4}$。分别测试不同随机算法在数据集CINA和a9a上的表现。我们采用网格搜索方法来确定随机算法中的参数值，对每个参数重复5次数值实验并取其平均表现。数值稳定参数均设置为$\epsilon = 10^{-7}$。


sgd without momentum
sgd with momentum
Adagrad
RMSProp
AdaDelta
Adam
函数值
epoch

Outline

Feedforward network

different notation: the variable is θ and x is the data.

- Given an input $a_0 = x$, the output $h(x, \theta) = a_L \in R^m$ can be obtained through a series of L layers as follows:

$$s_l = W_l a_{l-1}, \quad a_l = \phi_l(s_l), \qquad l = 1, 2, \ldots, L,$$

where ϕ_l is element-wise, and W_l is the weight in i-th layer .

- The variable: $\theta = [\mathrm{vec}(W_1)^\top \, \mathrm{vec}(W_2)^\top \ldots \mathrm{vec}(W_L)^\top]^\top$.
- Gradient by back-propagation Process:

$$g_l \leftarrow \mathcal{D}a_l \odot \phi_l'(s_l), \mathcal{D}W_l \leftarrow g_l a_{l-1}^\top, \mathcal{D}a_{l-1} \leftarrow W_l^\top g_l$$

- For convolution layer, the gradient can also be represented

$$\mathcal{D}W_l = G_l A_l^\top,$$

where G_l and A_l are matrices.

KL Divergence Objectives

- $Q_{x,y}$: the true data distribution.
- $\hat{Q}_{x,y}$: the training distribution given $\{(x_i, y_i)\}$
- $P_{x,y}(\theta)$: the learned distribution
- KL divergence: $KL(Q_{x,y} \| P_{x,y}) = \int q(x,y) \log \frac{q(x,y)}{p(x,y)} dxdy$.
- Goal: minimize the KL divergence from $\hat{Q}_{x,y}$ to $P_{x,y}(\theta)$

$$\mathbf{E}_{\hat{Q}_x}[KL(Q_{y|x} \| P_{y|x}(\theta))] = -\frac{1}{N} \sum_i \log p(y_i | h(x_i, \theta)).$$

Hence, our loss function is the negative log probability.

Kronecker product

- $A \otimes B$ denotes the Kronecker product between A and B:

$$A \otimes B \equiv \begin{bmatrix} [A]_{1,1}B & \cdots & [A]_{1,n}B \\ \vdots & \ddots & \vdots \\ [A]_{m,1}B & \cdots & [A]_{m,n}B \end{bmatrix}.$$

- $\mathrm{vec}(uv^\top) = v \otimes u$.
- $(A \otimes B)^{-1} = A^{-1} \otimes B^{-1}$.
- $(B^\top \otimes A)\,\mathrm{vec}(X) = \mathrm{vec}(AXB)$
- $\mathrm{vec}(G_i A_i^\top) = (A_i \otimes G_i)\,\mathrm{vec}(I)$.
- $(A \otimes B)(C \otimes D) = (AC) \otimes (BD)$ for any A, B, C, D with correct sizes.

Empirical Fisher Information Matrix (EFIM)

- Fisher Information Matrix

$$F = \mathbf{E}_{P_{x,y}}[\nabla\psi(h(x,\theta),y)\nabla\psi(h(x,\theta),y)^\top] = -\mathbf{E}_{P_{x,y}}[\nabla^2 \log p(y|h(x,\theta))]$$

- The EFIM is defined as follows:

$$\mathbf{F}(\theta) = \mathbf{E}_{\hat{Q}_{x,y}}\left[\nabla\psi(h(x,\theta),y)\nabla\psi(h(x,\theta),y)^\top\right]$$
$$= \begin{bmatrix} \mathbf{E}_{\hat{Q}_{x,y}}[a_0 a_0^\top \otimes g_1 g_1^\top] & \cdots & \mathbf{E}_{\hat{Q}_{x,y}}[a_0 a_{L-1}^\top \otimes g_1 g_L^\top] \\ \vdots & \ddots & \vdots \\ \mathbf{E}_{\hat{Q}_{x,y}}[a_{L-1} a_0^\top \otimes g_L g_1^\top] & \cdots & \mathbf{E}_{\hat{Q}_{x,y}}[a_{L-1} a_{L-1}^\top \otimes g_L g_L^\top] \end{bmatrix}$$

- The second-order Taylor approximation to KL divergence is the Fisher information matrix.
- KL divergence is an intrinsic dissimilarity measure on distributions: it doesn't care how the distributions are parameterized.

The Hessian Matrix

- The Hessian matrix is:

$$\mathbf{H}(\theta) = \mathbf{E}_{\hat{Q}_{x,y}}[\Sigma(\theta)] + \begin{bmatrix} \mathbf{E}_{\hat{Q}_{x,y}}[a_0 a_0^\top \otimes G_{1,1}] & \cdots & \mathbf{E}_{\hat{Q}_{x,y}}[a_0 a_{L-1}^\top \otimes G_{1,L}] \\ \vdots & \ddots & \vdots \\ \mathbf{E}_{\hat{Q}_{x,y}}[a_{L-1} a_0^\top \otimes G_{L,1}] & \cdots & \mathbf{E}_{\hat{Q}_{x,y}}[a_{L-1} a_{L-1}^\top \otimes G_{L,L}] \end{bmatrix}$$

  where

$$G_{ij} = \frac{\partial^2 \psi}{\partial s_i \partial s_j}, \quad \Sigma_{i,j} = \sum_p \frac{\partial^2 (s_j)_p}{\partial \operatorname{vec}(W_i) \partial \operatorname{vec}(W_j)} \odot (g_j)_p$$

- Note that $\Sigma_{ii} = 0$ for all $i = 1, \ldots, L$.
- Let θ^* be a global minimum. For θ in a sufficiently small neighborhood of θ^* and sufficiently large N, it holds with probability $1 - \delta$:

$$\|\mathbf{H}(\theta) - \mathbf{F}(\theta)\| < \epsilon$$

Natural Gradient Method

- The scheme:
$$\theta^{k+1} = \theta^k - \alpha^k \mathbf{F}(\theta^k)^{-1} g^k$$
- It holds $KL(P_{x,y}(\theta + d) \| P_{x,y}(\theta)) \to \frac{1}{2} d^\top F d$ as d goes to zero
- The steepest descent direction in the space of distributions where distance is (approximately) measured in local neighborhoods by the KL divergence:

$$-\sqrt{2}\frac{F^{-1}\nabla\Psi}{\|\nabla\Psi\|_{F^{-1}}} = \lim_{\epsilon\to 0}\frac{1}{\epsilon} \operatorname*{argmin}_{d\,:\,KL(P_{x,y}(\theta+d)\|P_{x,y}(\theta))\le\epsilon^2} \Psi(\theta + d).$$

- Similar to Gauss-Newton methods in nonlinear least squares?

Kronecker-factored Approximation to EFIM

- Block-diagonal (Layer-wise) Approximation to EFIM:

$$B = \mathrm{diag}\{F_1, \ldots, F_L\},$$

where F_l corresponds to the l-th layer.

- Note that $\mathcal{D}W_l = G_l A_l^\top$ and $\mathrm{vec}(G_l A_l^\top) = (A_l \otimes G_l)\,\mathrm{vec}(I)$. We have:

$$\begin{aligned}
F_l &= \mathbf{E}_{\hat{Q}_{x,y}}\left[\mathrm{vec}(\mathcal{D}W_l)\,\mathrm{vec}(\mathcal{D}W_l)^\top\right] \\
&= \mathbf{E}_{\hat{Q}_{x,y}}\left[(A_l \otimes G_l)\,\mathrm{vec}(I)\,\mathrm{vec}(I)^\top (A_l^\top \otimes G_l^\top)\right] \\
&\approx \mathbf{E}_{\hat{Q}_{x,y}}\left[(A_l \otimes G_l)(A_l^\top \otimes G_l^\top)\right] \\
&= \mathbf{E}_{\hat{Q}_{x,y}}\left[(A_l A_l^\top) \otimes (G_l G_l^\top)\right] \\
&\approx \mathbf{E}_{\hat{Q}_{x,y}}\left[A_l A_l^\top\right] \otimes \mathbf{E}_{\hat{Q}_{x,y}}\left[G_l G_l^\top\right] = \widehat{A} \otimes \widehat{G}
\end{aligned}$$

KFAC (James Martens and Roger Grosse)

- Delayed update of EFIM

$$\widehat{F}_t = (\widehat{A}^t_{\mathcal{B}^t} + \sqrt{\lambda} I) \otimes (\widehat{G}^t_{\mathcal{B}^t} + \sqrt{\lambda} I)$$

- Update the iteration ($g_{\mathcal{B}^k} = \mathrm{vec}(\mathcal{G}_{\mathcal{B}^k})$, $\Theta = \mathrm{vec}(\theta)$):

$$\theta^{k+1} = \theta^k - \alpha^k \widehat{F}_t^{-1} g_{\mathcal{B}^k},$$

or equivalently

$$\Theta^{k+1} = \Theta^k - \alpha^k (\widehat{G}^t_{\mathcal{B}^t} + \sqrt{\lambda} I)^{-1} \mathcal{G}_{\mathcal{B}^k} (\widehat{A}^t_{\mathcal{B}^t} + \sqrt{\lambda} I)^{-1}.$$

- Use the momentum technique to generate direction.
- Improvement: block diagonal approximation to $\widehat{A}^t_{\mathcal{B}^t}$ and $\widehat{G}^t_{\mathcal{B}^t}$