Optimization for Machine Learning
机器学习中的优化方法

陈 程

华东师范大学 软件工程学院

chchen@sei.ecnu.edu.cn

Outline

Review

Stochastic optimization problem:

$$\min_{\mathbf{x}\in\mathbb{R}^d} F(\mathbf{x}) \triangleq \underbrace{\mathbb{E}_{\xi}[f(\mathbf{x};\xi)]}_{\text{expectation setting}},$$

where the random variable $\xi \sim \mathcal{D}$.

The finite-sum setting is a special case of the expectation setting:

$$F(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{x}).$$

Clear up

Stochastic gradient descent:

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \eta_t g(\mathbf{x}_t, \xi) \quad \text{(expectation setting)}$$

Suppose we return a weighted average

$$\tilde{\mathbf{x}}_t = \sum_{k=0}^{t} \frac{\eta_k}{\sum_{j=0}^{t} \eta_j} \mathbf{x}_k$$

If F is convex, we have

$$\mathbb{E}[F(\tilde{\mathbf{x}}_t) - F(\mathbf{x}^*)] \leq \frac{\|\mathbf{x}_0 - \mathbf{x}^*\|_2^2 + \sum_{k=0}^{t} \sigma^2 \eta_k^2}{2 \sum_{k=0}^{t} \eta_k}.$$

Review

Stochastic gradient descent:

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \eta_t \nabla f_{i_t}(\mathbf{x}_t) \quad \text{(finite-sum setting)}$$

For fixed step size, SGD achieves

$$\mathbb{E}\left[\|\mathbf{x}_t - \mathbf{x}^*\|_2^2\right] \leq (1 - 2\eta\mu)^t \|\mathbf{x}_0 - \mathbf{x}^*\|_2^2 + \frac{\eta\sigma^2}{2\mu}.$$

How to reduce the variance of the gradient estimator?

Outline

1. Review
2. Stochastic variance reduced gradient
3. Nonconvex optimization
4. Stochastic nonconvex optimization

Stochastic variance reduced gradient (SVRG)

NOTE: For some $\mathbf{v}_t$ with $\mathbb{E}[\mathbf{v}_t] = 0$, $\mathbf{g}_t = \nabla f_{i_t}(\mathbf{x}_t) - \mathbf{v}_t$ is still an unbiased estimator of $\nabla F(\mathbf{x}_t)$.

If we have access to a history point $\tilde{\mathbf{x}}$ and $\nabla F(\tilde{\mathbf{x}})$, how to build a unbiased gradient estimator with converges to $\mathbf{0}$?

$$\underbrace{\nabla f_i(\mathbf{x}_t) - \nabla f_i(\tilde{\mathbf{x}})}_{\to \mathbf{0} \text{ if } \mathbf{x}_t \approx \tilde{\mathbf{x}}} + \underbrace{\nabla F(\tilde{\mathbf{x}})}_{\to \mathbf{0} \text{ if } \tilde{\mathbf{x}} \approx \mathbf{x}^*}$$

where i is randomly sampled from $\{1, \ldots, n\}$.

- an unbiased estimator of $\nabla F(\mathbf{x}_t)$
- converges to $\mathbf{0}$ if $\mathbf{x}_t \approx \tilde{\mathbf{x}} \approx \mathbf{x}^*$

Stochastic variance reduced gradient (SVRG)

- operate in epochs
- in the s-th epoch
  - **beginning:** take a snapshot $\tilde{\mathbf{x}}$ of the current iterate, and compute the batch gradient

$$\nabla F(\tilde{\mathbf{x}}) = \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(\tilde{\mathbf{x}}).$$

  - **inner loop:** use the snapshot point to help reduce variance

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \eta_t(\nabla f_i(\mathbf{x}_t) - \nabla f_i(\tilde{\mathbf{x}}) + \nabla F(\tilde{\mathbf{x}})),$$

Stochastic variance reduced gradient (SVRG)

Algorithm 1 Stochastic Variance Reduced Gradient

Input: $\mathbf{x}_0$, η, m, S
$\tilde{\mathbf{x}}^{(0)} = \mathbf{x}_0$
for $s = 0, \ldots, S-1$
  $\mathbf{x}_0 = \tilde{\mathbf{x}}^{(s)}$
  for $t = 0, \ldots, m-1$
    draw i_t from $\{1, \ldots, n\}$ uniformly at random
    $\mathbf{x}_{t+1} = \mathbf{x}_t - \eta(\nabla f_{i_t}(\mathbf{x}_t) - \nabla f_{i_t}(\tilde{\mathbf{x}}^{(s)}) + \nabla F(\tilde{\mathbf{x}}^{(s)})),$
  end for
  Option I: $\tilde{\mathbf{x}}^{(s+1)} = \mathbf{x}_m$
  Option II: $\tilde{\mathbf{x}}^{(s+1)} = \mathbf{x}_t$ for randomly chosen $t \in \{0, \ldots, m-1\}$
end for
Output: $\tilde{\mathbf{x}}^{(S)}$

Remark

- constant stepsize η
- each epoch contains $2m + n$ gradient computations
- the average per-iteration cost of SVRG is comparable to that of SGD if $m \gtrsim n$

Convergence analysis

Suppose $F(\mathbf{x})$ is L-smooth and μ-strongly convex. Let $\eta = \Theta(1/L)$ and $m = \Theta(\kappa)$ is sufficient large so that

$$\rho = \frac{1}{\mu\eta(1-2L\eta)m} + \frac{2L\eta}{1-2L\eta} < 1,$$

then SVRG holds that

$$\mathbb{E}\big[F(\tilde{\mathbf{x}}^{(s)}) - F(\mathbf{x}^*)\big] \le \rho^s(F(\mathbf{x}_0) - F(\mathbf{x}^*)).$$

To achieve

$$\mathbb{E}\big[F(\tilde{\mathbf{x}}^{(s)}) - F(\mathbf{x}^*)\big] \le \epsilon$$

we only require at most $\mathcal{O}((n+\kappa)\log(1/\epsilon))$ number of gradient computations.

Convergence analysis

Important Lemma:

$$\mathbb{E}_t \left[\left\| \mathbf{g}_t^{(s)} \right\|_2^2 \right] \leq 4L \left[F(\mathbf{x}_t^{(s)}) - F(\mathbf{x}^*) + F(\tilde{\mathbf{x}}^{(s)}) - F(\mathbf{x}^*) \right]$$

Summary

$$\min_{\mathbf{x}\in\mathbb{R}^d} F(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{x}).$$

	iteration complexity	per-iteration	total
batch GD	$\kappa \log(1/\epsilon)$	n	$n\kappa \log(1/\epsilon)$
SGD	$1/\epsilon$	1	$1/\epsilon$
SVRG	$\log(1/\epsilon)$	$n+\kappa$	$(n+\kappa)\log(1/\epsilon)$

Table: Convergence rate for the strongly convex case

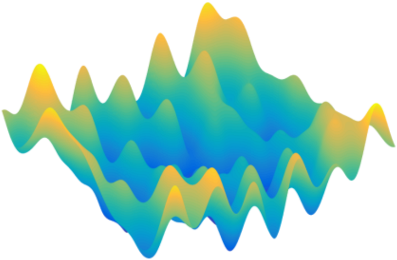
Outline

Nonconvex problems

Many objective functions in machine learning are nonconvex:

- low-rank matrix completion
- mixture models
- learning deep neural nets
- ...

Challenges



- there may be local minima everywhere
- no algorithm can solve nonconvex problems efficiently in all cases

Typical convergence guarantees

We cannot hope for efficient global convergence to global minima in general, but we may have

- convergence to stationary points ,i.e., $\nabla f(\mathbf{x}) = 0$
- convergence to local minima
- local convergence to global minima i.e., when initialized suitably

Making gradients small

Suppose we aim to find a stationary point, which means that our goal is merely to find a point $\mathbf{x}$ with

$$\|\nabla f(\mathbf{x})\|_2 \leq \epsilon \text{ (called } \epsilon\text{-approximate stationary point)}$$

ϵ-approximate stationary point does not imply local minima for nonconvex optimization.

Making gradients small

Let f be L-smooth and $\eta_t = \eta = \frac{1}{L}$, then GD obeys

$$\min_{0 \le k < t} \|\nabla f(\mathbf{x}_t)\|_2 \le \sqrt{\frac{2L(f(\mathbf{x}_0) - f(\mathbf{x}^*))}{t}}.$$

- GD finds an ϵ-approximate stationary point in $O(1/\epsilon^2)$ iterations.
- does not imply GD converges to stationary points; it only says that there exists an approximate stationary point in the GD trajectory

Outline

Stochastic nonconvex optimization

Stochastic nonconvex optimization:

$$\min_{\mathbf{x} \in \mathbb{R}^d} F(\mathbf{x}) \triangleq \mathbb{E}_\xi[f(\mathbf{x};\xi)],$$

where $f(\mathbf{x};\xi)$ is L-smooth and potentially nonconvex.

Our goal is to find a first-order stationary point $\mathbf{x}$ such that

$$\mathbb{E}[\|\nabla F(\mathbf{x})\|_2] \leq \epsilon.$$

Assumption:

$$\mathbb{E}_\xi[\|f(\mathbf{x},\xi) - F(\mathbf{x})\|_2^2] \leq \sigma^2.$$

SGD for nonconvex optimization

Stochastic gradient descent:

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \eta_t \nabla f(\mathbf{x}_t, \xi_t).$$

- Return $\bar{\mathbf{x}}$ chosen uniformly at random from $\{\mathbf{x}_0, \ldots, \mathbf{x}_{t-1}\}$.
- If we choose

$$\eta = \eta_t = \frac{1}{L} \min\left\{\frac{\epsilon^2}{2\sigma^2}, 1\right\} \text{ and } t = \frac{4(F(\mathbf{x}_0) - F(\mathbf{x}^*))}{\epsilon^2 \eta},$$

then

$$\mathbb{E}[\|\nabla F(\bar{\mathbf{x}})\|_2] \leq \epsilon.$$