Stochastic gradient descent

Olivier Fercoq

Télécom Paris

Setup

- We consider a function

$$f : \mathbb{R}^d \times \Xi \to \mathbb{R}$$
$$(x, t) \mapsto f(x, t)$$

and a random variable $\xi : \Omega \to \Xi$ with law $\mathcal{D}$.

the average of functions

- We would like to find

$$x^* \in \arg\min_{x \in \mathbb{R}^d} \mathbb{E}_{\xi \sim \mathcal{D}}[f(x, \xi)]$$

- Challenge: the law $\mathcal{D}$ is unknown. Yet, we can sample $\xi_i \sim \mathcal{D}$.

$$x^* \in \arg\min \frac{1}{n} \sum f_i(x, \xi)$$

Gradient descent

- Empirical risk minimization

$$F(x) = \mathbb{E}_{\xi\sim\mathcal{D}}[f(x,\xi)] \approx F_N(x) = \frac{1}{N}\sum_{i=1}^{N} f(x,\xi_i)$$

- We are looking for

$$x_N^* \in \arg\min_{x\in\mathbb{R}^d} F_N(x)$$

- Algorithm

$$x^0 \in \mathbb{R}^d$$

$$x^{(k+1)} = x^{(k)} - \gamma \nabla F_N(x) = x^{(k)} - \frac{\gamma}{N}\sum_{i=1}^{N} \nabla_x f(x,\xi_i).$$

$\gamma \in]0, 2/L[$: step size

- Dimension of $x = d$, number of samples $= N$
  Cost per iteration: $O(dN)$

Stochastic gradient descent

- When N is large, a cost per iteration of $O(Nd)$ is a lot
- Idea: update x^k each time we get a sample
- Algorithm

$x^{(0)} \in \mathbb{R}^d$

$\xi_{k+1} \sim D$

$x^{(k+1)} = x^{(k)} - \gamma_k \nabla f(x_k, \xi_{k+1})$

$\nabla f(x_k, \xi_{k+1})$ is an unbiased estimator

$$F(x_k) = E_{\xi \sim D}[f(x_k, \xi)]$$

$$\nabla F(x_k) = E_{\xi \sim D}[\nabla f(x_k, \xi)]$$

$$= E_{\xi_{k+1} \sim D}[\nabla f(x_k, \xi_{k+1})]$$

ξ_{k+1} is independent of x_k and $\xi_{k+1} \sim D$

- Cost per iteration $O(d)$

Stochastic gradient descent

- When N is large, a cost per iteration of $O(Nd)$ is a lot
- Idea: update x^k each time we get a sample
- Algorithm

$$
\begin{aligned}
&x^0 \in \mathbb{R}^d \\
&\forall k \in \mathbb{N}: \\
&\qquad \xi_{k+1} \sim \mathcal{D} \\
&\qquad x^{k+1} = x^k - \gamma_k \nabla f(x^k, \xi_{k+1})
\end{aligned}
$$

每次只用一个 Sample

- Cost per iteration $O(d)$

Converge of stochastic gradient descent

Theorem

Suppose that:

- $(x \mapsto f(x,\xi))$ *is convex and differentiable for all* ξ*,*
- *there exists* $C > 0$ *such that* $\mathbb{E}(\|\nabla f(x,\xi)\|^2) \leq C$ *for all* x
- *there exists* $x^* \in \arg\min F$*,*
- *the sequence* γ_k *is deterministic.*

Then the iterates of SGD $x_{k+1} = x_k - \gamma_k \nabla f(x_k, \xi_{k+1})$ *satisfy*

$$\mathbb{E}\Big[F(\bar{x}_k^\gamma) - F(x^*)\Big] \leq \frac{\mathbb{E}[\|x_0 - x^*\|^2] + C\sum_{l=0}^k \gamma_l^2}{2\sum_{l=0}^k \gamma_l}$$

where $\bar{x}_k^\gamma = \dfrac{\sum_{l=0}^k \gamma_l x_l}{\sum_{j=0}^k \gamma_j}$*.*

Proof 1/2

$$\mathbb{E}\left[F(\bar{x}_k^\gamma) - F(x^*)\right] \leq \frac{\mathbb{E}[\|x_0 - x^*\|^2] + C\sum_{l=0}^k \gamma_l^2}{2\sum_{l=0}^k \gamma_l}$$

Proof 2/2

$$\mathbb{E}\left[F(\bar{x}_k^{\gamma}) - F(x^*)\right] \leq \frac{\mathbb{E}[\|x_0 - x^*\|^2] + C\sum_{l=0}^{k}\gamma_l^2}{2\sum_{l=0}^{k}\gamma_l}$$

Choice of the sequence (γ_k)

$$\mathbb{E}\left[F(\bar{x}_k^{\gamma}) - F(x^*)\right] \leq \frac{\mathbb{E}[\|x_0 - x^*\|^2] + C\sum_{l=0}^{k}\gamma_l^2}{2\sum_{l=0}^{k}\gamma_l}$$

- What choice of γ_k ensures a faster decrease?
- We consider $\gamma_k = \dfrac{\gamma_0}{(k+1)^{\alpha}}$ for a given $\alpha > 0$.

	$\dfrac{1}{\sum_{j=0}^{k}\gamma_j}$	$\dfrac{\sum_{l=1}^{k}\gamma_l^2}{\sum_{j=1}^{k}\gamma_j}$	
$0 < \alpha < 1/2$	$O\left(\dfrac{1}{k^{1-\alpha}}\right)$	$O\left(\dfrac{1}{k^{\alpha}}\right)$	
$\alpha = 1/2$	$O\left(\dfrac{1}{k^{1/2}}\right)$	$O\left(\dfrac{\ln(k)}{k^{1/2}}\right)$	$\rightarrow$ best
$1/2 < \alpha < 1$	$O\left(\dfrac{1}{k^{1-\alpha}}\right)$	$O\left(\dfrac{1}{k^{1-\alpha}}\right)$	

- The best rate is $O\left(\dfrac{\ln(k)}{k^{1/2}}\right)$ for $\gamma_k = \frac{\gamma_0}{\sqrt{k+1}}$

Whiteboard

Case where we know the number of iteration in advance

$$\mathbb{E}\left[F(\bar{x}_k^\gamma) - F(x^*)\right] \leq \frac{\mathbb{E}[\|x_0 - x^*\|^2] + C\sum_{l=0}^{k}\gamma_l^2}{2\sum_{l=0}^{k}\gamma_l}$$

- We choose $\gamma_l = \frac{a}{\sqrt{k}}$
- $\sum_{l=0}^{k-1}\gamma_l = a\sqrt{k}$
- $\sum_{l=0}^{k-1}\gamma_l^2 = a^2$
- Constant step size: the algorithm does not converge
- Yet, for iteration k

$$\mathbb{E}\left[F(\bar{x}_{k-1}^\gamma) - F(x^*)\right] \leq \frac{\mathbb{E}[\|x_0 - x^*\|^2] + Ca^2}{2a\sqrt{k}}$$

Extension

Setup

$$\min_x \mathbb{E}_{\xi \sim \mathcal{D}}[f(x, \xi)] + g(x)$$

where g is convex, not differentiable but has a simple proximal operator

Proximal stochastic gradient descent

$$\begin{aligned}
&x^0 \in \mathbb{R}^d \\
&\forall k \in \mathbb{N}: \\
&\qquad \xi_{k+1} \sim \mathcal{D} \\
&\qquad x^{k+1} = \mathrm{prox}_{\gamma_k g}\left(x_k - \gamma_k \nabla f(x_k, \xi_{k+1})\right)
\end{aligned}$$

Optimization and statistics

- Ideal estimator

$$x^* \in \operatorname{argmin} E_{\xi \to D}\left[f(x, \xi)\right]$$

- Empirical risk minimization

$$x_N^* \in \operatorname*{argmin}_{x \in \mathbb{R}^d} \frac{1}{N} \sum_{i=1}^{N} f(x, \xi_i)$$

- What the algorithm returns

x_k that is close to x_N^*

Optimization and statistics

- Ideal estimator

$$x^* \in \arg\min_x F(x) = \mathbb{E}_{x\sim\mathcal{D}}[f(x,\xi)]$$

- Empirical risk minimization

$$x_N^* \in \arg\min_x F_N(x) = \frac{1}{N}\sum_{i=1}^{N} f(x,\xi_i)$$

- What the algorithm returns

$$\hat{x}_N = x^k$$

- $\mathbb{E}[F(x_k) - F(x^*)] =$

$$\underbrace{\mathbb{E}[F_N(x_k) - F_N(x_N^*)]}_{\text{optimisation error } \mathcal{E}_{\text{opt}}} + \underbrace{\mathbb{E}[F_N(x_N^*) - F(x^*)] + \mathbb{E}[F(x_k) - F_N(x_k)]}_{\text{estimation error } \mathcal{E}_{\text{est}}}$$

Estimation/optimization tradeoff

$$
\begin{aligned}
&\mathbb{E}[F(x_k) - F(x^*)] \\
&\quad = \mathbb{E}[F_N(x_k) - F_N(x_N^*)] + \mathbb{E}[F_N(x_N^*) - F(x^*)] + \mathbb{E}[F(x_k) - F_N(x_k)] \\
&\quad = \mathcal{E}_{\text{opt}} + \mathcal{E}_{\text{est}}
\end{aligned}
$$

estimation error	$\mathcal{E}_{\text{est}} \leq c\sqrt{\frac{d}{N}}$	
step size	gradient descent $\gamma = 1/L$	stochastic gradient $\gamma = \frac{a}{\sqrt{k}}$
optimization error after k iterations	$\mathcal{E}_{\text{opt}} \leq \frac{C_1}{k}$	$\mathcal{E}_{\text{opt}} \leq \frac{C_2}{\sqrt{k}}$
cost for 1 iteration	Nd	d
total cost for $\mathcal{E}_{\text{opt}} \approx \mathcal{E}_{\text{est}}$	$C_3 Nd\sqrt{\frac{N}{d}}$	$C_4 d(\sqrt{\frac{N}{d}})^2$

Estimation/optimization tradeoff

$$
\begin{aligned}
&\mathbb{E}[F(x_k) - F(x^*)] \\
&\quad = \mathbb{E}[F_N(x_k) - F_N(x_N^*)] + \mathbb{E}[F_N(x_N^*) - F(x^*)] + \mathbb{E}[F(x_k) - F_N(x_k)] \\
&\quad = \mathcal{E}_{\text{opt}} + \mathcal{E}_{\text{est}}
\end{aligned}
$$

estimation error	$\mathcal{E}_{\text{est}} \leq c\sqrt{\frac{d}{N}}$	
step size	gradient descent $0 < \gamma < \frac{2}{L}$	stochastic gradient $\gamma = \frac{a}{\sqrt{k}}$
optimization error after k iterations	$\leq \frac{c}{k}$	$\leq \frac{c'}{\sqrt{k}}$
cost for 1 iteration	dN	d
total cost for $\mathcal{E}_{\text{opt}} \approx \mathcal{E}_{\text{est}}$	$O(N^{3/2} d^{1/2})$	$O(N)$

- $O(N\sqrt{N})$ vs $O(N)$
- Choosing (γ_k) is not easy and problem-dependent