Optimization for Machine Learning

Xiao Wang

Shanghai University of Finance and Economics

May 7, 2021

Stochastic Gradient Descent

Definition

Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex function in some convex set $\mathcal{X}$. The SGD is given as

$$x_{k+1} = x_k - \alpha_k v_k,$$

where $\mathbb{E}[v_k|x_k] = \nabla f(x_k)$

- α_k is called the step-size.
- α_k must be vanishing s.t. SGD converges.
- v_k and x_k are random vectors.

Stochastic Gradient Descent

Theorem

Let $f : \mathbb{R}^d \to \mathbb{R}$ be μ-strongly convex. Assume that $\mathbb{E}[\|v_k\|^2] \leq \rho^2$. Let x^* be a minimizer. It holds for $\alpha_k = \frac{1}{\mu k}$,

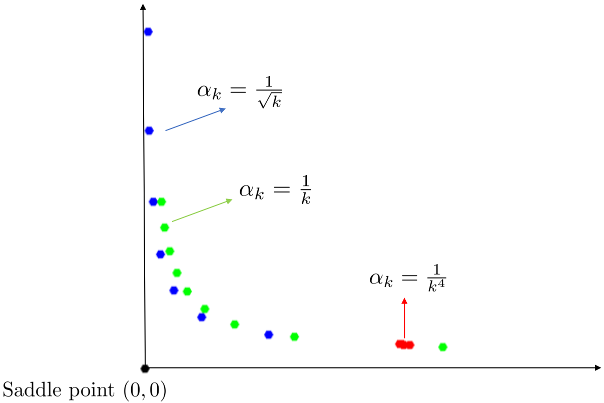
$$\mathbb{E}\left[f\left(\frac{1}{T}\sum_t x_t\right)\right] - f(x^*) \leq \frac{\rho^2}{2\mu T}(1 + \log T).$$

- α_k scales as $\frac{1}{k}$ and is vanishing.
- For $T = \Theta\left(\frac{1}{\epsilon}\log\frac{1}{\epsilon}\right)$ we get error ϵ.

Stochastic Gradient Descent

More on step-size

GD on $f(x,y) = x^2 - y^2$ $\Rightarrow$ $$\begin{aligned} x_{k+1} &= x_k - \alpha_k \cdot 2x_k \\ y_{k+1} &= y_k + \alpha_k \cdot 2y_k \end{aligned}$$



Stochastic Gradient Descent

Example: Coordinate descent

Let f be convex differentiable in some convex set $\mathcal{X}$. Coordinate Descent is defined:

$$x_{t+1} = x_t - \alpha_t \frac{\partial f}{\partial x_i} e_i$$

for iteratively chosen $i \in [d]$.

- Similar guarantees with GD as long as each coordinate is taken often.
- If coordinate i is chosen uniformly at random, then $\mathbb{E}\left[\frac{\partial f}{\partial x_i}\right] = \frac{1}{n}\nabla f(x)$.
- Open question: Does deterministic (block) coordinate descent almost always avoid saddle points with vanishing step-size?

Stochastic Gradient Descent

Risk Minimization

Let $\ell(x, z) : \mathcal{X} \times \mathcal{Z} \to \mathbb{R}$ be a risk function and D osme unkown distribution we can get samples from. We are interested in solving:

$$\min_{x \in \mathcal{X}} L(x)$$

where $L(x) = \mathbb{E}_{z \sim D}[\ell(x, z)]$.

Stochastic Gradient Descent

Risk Minimization

Let $\ell(x, z) : \mathcal{X} \times \mathcal{Z} \to \mathbb{R}$ be a risk function and D osme unkown distribution we can get samples from. We are interested in solving:

$$\min_{x \in \mathcal{X}} L(x)$$

where $L(x) = \mathbb{E}_{z \sim D}[\ell(x, z)]$.

Question:

Connection to optimization for neural networks?

Stochastic Gradient Descent

Risk Minimization

Let $\ell(x, z) : \mathcal{X} \times \mathcal{Z} \to \mathbb{R}$ be a risk function and D osme unkown distribution we can get samples from. We are interested in solving:

$$\min_{x \in \mathcal{X}} L(x)$$

where $L(x) = \mathbb{E}_{z \sim D}[\ell(x, z)]$.

Approach one:

- Take enough samples z_i independently and consider the estimate $\bar{L}(x) = \frac{1}{n} \sum_i \ell(x, z_i)$. (Law of Large Numbers)
- Run first order optimization algorithm on $\bar{L}(x)$ to minimize it.

Stochastic Gradient Descent

Risk Minimization

Let $\ell(x, z) : \mathcal{X} \times \mathcal{Z} \to \mathbb{R}$ be a risk function and D osme unkown distribution we can get samples from. We are interested in solving:

$$\min_{x \in \mathcal{X}} L(x)$$

where $L(x) = \mathbb{E}_{z \sim D}[\ell(x, z)]$.

Approach two: SGD

- For each iteration $t + 1$, take a fresh sample z_t independently from $z_1, ..., z_{t-1}$ and consider the unbiased estimate $\nabla_x \ell(x_t, z_t)$.
- Update $x_{t+1} = x_t - \alpha_t \nabla_x \ell(x_t, z_t)$.
- Return for $\frac{1}{T} \sum x_t$.

Stochastic Gradient Descent

Question:

Why SGD works well even in non-convex settings? (Converges to global minima, not stuck at saddle point etc)

Langevin Equation and Sampling

Question: How to generate random samples from $\mathbb{R}^d$ such that these points satisfies certain probability distribution?

Langevin Equation and Sampling

Question: How to generate random samples from $\mathbb{R}^d$ such that these points satisfies certain probability distribution?

Langevin equation

$$dX_t = -\nabla f(X_t)dt + \sqrt{2}dB_t$$

Langevin Equation and Sampling

Question: How to generate random samples from $\mathbb{R}^d$ such that these points satisfies certain probability distribution?

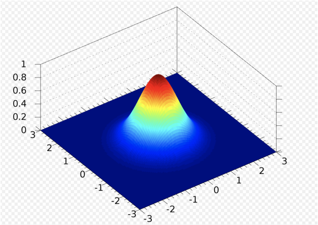
Langevin equation

$$dX_t = -\nabla f(X_t)dt + \sqrt{2}dB_t$$

As a random variable, X_t has its density function, denoted by ρ_t, and this density function evolves as X_t evolves according the stochastic differential equaiton.

Log-concave distributions

- Function $p(x)$ is called log-concave if $logp(tx + (1-t)y) \geq t \log p(x) + (1-t) \log p(y)$ for all $0 \leq t \leq 1$, or simply, $\log p(x)$ is concave.
- Distribution whose density function is log-concave is called log-concave distribution.
- Example: Gaussian distribution, density function $p(x) = e^{-\|x\|^2}$.



Sampling by Langevin Dynamics

- To sample from a distribution $\nu \propto e^{-f(x)}$ on $\mathbb{R}^d$, we often use the Langevin algorithm:
$$x_{t+1} = x_t - \alpha \nabla f(x_t) + \sqrt{2\alpha} z_0$$
where z_0 is the Gaussian noise.
- This algorithm is expected to converge to a biased distribution that is close to ν.
- For case of log-concave distribution, there are exitensively amout research, the convergence is rapid.

Sampling vs. Optimization

Informally:

- Optimization is sampling from a Dirac distribution.
- Sampling is optimization in the space of distributions.

Sampling vs. Optimization

Informally:

- Optimization is sampling from a Dirac distribution.
- Sampling is optimization in the space of distributions.

Recall the Langevin equation

$$dX_t = -\nabla f(X_t)dt + \sqrt{2}dB_t$$

The density function of X_t satisfies

$$\frac{\partial p(x,t)}{\partial t} = \nabla \cdot (p(x,t)\nabla f(x)) + \Delta p(x,t).$$

Reading: Fokker-Planck equation.

Readings

- Sampling can be faster than optimization. Yi-an Ma, Yuansi Chen, Chi Jin, Nicolas Flammarion, and Michael I. Jordan. 2019.
- Dynamical, symplectic and stochastic perspectives on gradient-based optimization. Michael I. Jordan. 2018.

Thank You!