Finite Sample Convergence Rates of Zero-Order Stochastic Optimization Methods

John C. Duchi **Michael I. Jordan** **Martin J. Wainwright** **Andre Wibisono**

{jduchi,jordan,wainwrig,wibisono}@eecs.berkeley.edu

Introduction and Problem Statement

Setup: Want to solve the following stochastic convex optimization problem:

$$\min_{\theta\in\Theta} f(\theta) := \mathbb{E}_P[F(\theta; X)] = \int_{\mathcal{X}} F(\theta; x)\, dP(x)$$

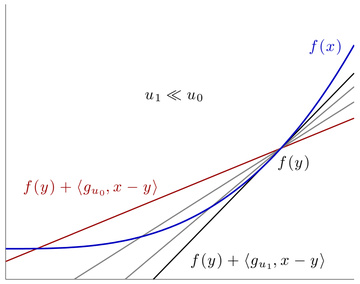
where $\Theta \subseteq \mathbb{R}^d$ is a compact convex set and $F(\cdot; x)\colon \Theta \to \mathbb{R}$ is closed convex.

- Assume evaluating (sub)gradient $\nabla F(\cdot; x)$ is difficult, or expensive.
- Can only obtain function evaluations $F(\cdot; X)$ for samples $X \sim P$.

Zero-order optimization:
Approximate gradient by difference of function values.
E.g. for $d = 1$ and u small,

$$f'(y) \approx g_u := \frac{f(y+u) - f(y)}{u}$$

Goal: Analyze upper and lower bounds on rate of convergence.



Stochastic Mirror Descent

Stochastic optimization methods for when we can calculate subgradients.

- Choose a 1-strongly convex proximal function $\psi\colon \Theta \to \mathbb{R}$.
- This defines a Bregman divergence $D_\psi\colon \Theta \times \Theta \to \mathbb{R}_+$ via

$$D_\psi(\theta, \tau) := \psi(\theta) - \psi(\tau) - \langle \nabla\psi(\tau), \theta - \tau\rangle \geq \frac{1}{2}\|\theta - \tau\|^2.$$

Algorithm: Sample $X^t \sim P$, calculate subgradient $g^t = \mathsf{g}(\theta^t; X^t)$, and set

$$\theta^{t+1} = \operatorname*{argmin}_{\theta\in\Theta} \left\{ \langle g^t, \theta\rangle + \frac{1}{\alpha(t)} D_\psi(\theta, \theta^t) \right\}. \tag{1}$$

Convergence rate: If $D_\psi(\theta^*, \theta) \leq \frac{1}{2}R^2$ and $\mathbb{E}[\|\mathsf{g}(\theta; X)\|_*^2] \leq G^2$ for $\theta \in \Theta$, then with stepsize $\alpha(t) = \alpha R/G\sqrt{t}$, we have

$$\mathbb{E}[f(\widehat{\theta}(k))] - f(\theta^*) \leq 2\frac{RG}{\sqrt{k}} \max\{\alpha, \alpha^{-1}\}$$

where $\widehat{\theta}(k) = \frac{1}{k}\sum_{t=1}^k \theta^t$ and $\theta^* \in \operatorname{argmin}_{\theta\in\Theta} f(\theta)$.

Derivative-Free Mirror Descent

Stochastic mirror descent with two-point gradient estimates.

Algorithm: At each iteration t:

- Sample $X^t \sim P$.
- Sample independent $Z^t \sim \mu$ with $\mathbb{E}_\mu[ZZ^\top] = I_d$.
- Set gradient estimator

$$g^t = \frac{F(\theta^t + u_t Z^t; X^t) - F(\theta^t; X^t)}{u_t} Z^t. \tag{2}$$

- Apply mirror descent update (1) to compute θ^{t+1}.

Upper Bounds on Convergence Rate

Intuition for two-point gradient estimates:

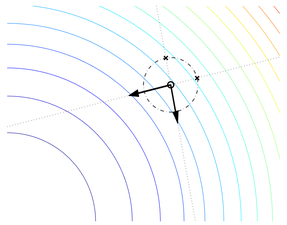
- At any point θ and direction z in Θ, for small $u > 0$ we have the approximation to the directional derivative:

$$\frac{f(\theta + uz) - f(\theta)}{u} \approx f'(\theta, z) := \lim_{h\downarrow 0} \frac{f(\theta + hz) - f(\theta)}{h}.$$

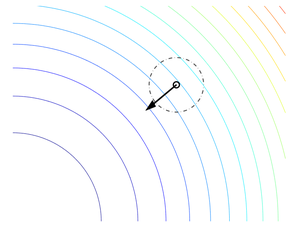
- If f is differentiable at θ, then $f'(\theta, z) = \langle \nabla f(\theta), z\rangle$.
- If the random vector Z satisfies $\mathbb{E}[ZZ^\top] = I_d$, then

$$\mathbb{E}[f'(\theta, Z)Z] = \mathbb{E}[\langle \nabla f(\theta), Z\rangle\, Z] = \mathbb{E}[ZZ^\top \nabla f(\theta)] = \nabla f(\theta).$$

- Thus, g^t in (2) is a nearly unbiased estimator of the gradient $\nabla f(\theta^t)$.



Estimates from random samples



Average is close to the true gradient

Assume:

- $F(\cdot; x)$ has $L(x)$-Lipschitz continuous gradient with $L^2 := \mathbb{E}[L(X)^2] < \infty$.
- There is a constant $s(d)$ such that $\mathbb{E}[\|\langle g, Z\rangle\, Z\|_*^2] \leq s(d)\, \|g\|_*^2$ for any $g \in \mathbb{R}^d$.

Convergence Rate: *With step and perturbation sizes*

$$\alpha(t) = \frac{R}{2G\sqrt{s(d)}} \cdot \frac{1}{\sqrt{t}} \quad \text{and} \quad u_t = \frac{u}{t},$$

the optimization error in derivative-free mirror descent is

$$\mathbb{E}[f(\widehat{\theta}(k))] - f(\theta^*) \leq 2\frac{RG\sqrt{s(d)}}{\sqrt{k}} + \mathcal{O}\left(u^2 \frac{\log k}{k}\right).$$

Examples

1. Stochastic gradient descent: if $\Theta \subseteq \{\theta \in \mathbb{R}^d\colon \|\theta\|_2 \leq R\}$, $\psi(\theta) = \frac{1}{2}\|\theta\|_2^2$, and $Z \sim \text{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$, then

$$\mathbb{E}[f(\widehat{\theta}(k))] - f(\theta^*) \leq 2\frac{RG\sqrt{d}}{\sqrt{k}} + \mathcal{O}\left(u^2 \frac{\log k}{k}\right).$$

2. High-dimensional version: if $\Theta \subseteq \{\theta \in \mathbb{R}^d\colon \|\theta\|_1 \leq R\}$, use proximal function $\psi(\theta) = \frac{1}{2(p-1)}\|\theta\|_p^2$ with $p = 1 + 1/\log d$, and $Z \sim \text{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$. Then

$$\mathbb{E}[f(\widehat{\theta}(k))] - f(\theta^*) \leq C\frac{RG\sqrt{d}\log d}{\sqrt{k}} + \mathcal{O}\left(u^2 \frac{\log k}{k}\right).$$

Lower Bounds on Convergence Rate

- Let $\mathbb{A}_k$ denote the set of methods that observe a sequence of data pairs $Y^t = (F(\theta^t, X^t),\ F(\tau^t, X^t))$, $1 \leq t \leq k$, and return an estimate $\widehat{\theta}(k) \in \Theta$.
- Let $\mathcal{F}_G$ denote the class of functions we want to optimize, where for each $(F, P) \in \mathcal{F}_G$ the subgradient $\mathsf{g}(\theta; X)$ satisfies $\mathbb{E}_P[\|\mathsf{g}(\theta; X)\|_*^2] \leq G^2$.
- For each $\mathcal{A} \in \mathbb{A}_k$ and $(F, P) \in \mathcal{F}_G$, consider the **optimization gap**:

$$\epsilon_k(\mathcal{A}, F, P, \Theta) := f(\widehat{\theta}(k)) - \inf_{\theta\in\Theta} f(\theta) = \mathbb{E}_P\left[F(\widehat{\theta}(k); X)\right] - \inf_{\theta\in\Theta} \mathbb{E}_P\left[F(\theta; X)\right].$$

- Define **minimax error** of zero-order optimization:

$$\epsilon_k^*(\mathcal{F}_G, \Theta) := \inf_{\mathcal{A}\in\mathbb{A}_k} \sup_{(F,P)\in\mathcal{F}_G} \mathbb{E}[\epsilon_k(\mathcal{A}, F, P, \Theta)].$$

Minimax Lower Bounds: If $\Theta = \{\theta \in \mathbb{R}^d\colon \|\theta\|_2 \leq R\}$ or $\Theta = \{\theta \in \mathbb{R}^d\colon \|\theta\|_1 \leq R\}$, then

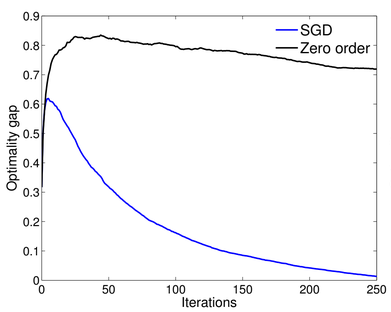
$$\epsilon_k^*(\mathcal{F}_G, \Theta) \geq c\frac{GR\sqrt{d}}{\sqrt{k}}.$$

- Our mirror descent-based algorithm achieves optimal convergence rates!
- $\sqrt{d}$ optimality gap between stochastic zero-order and gradient methods.

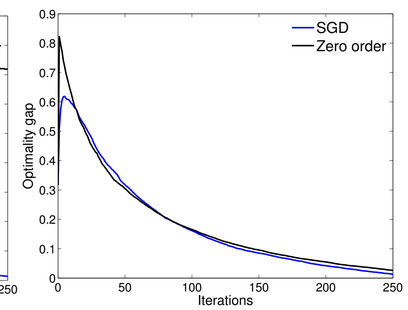
Simulation Results

Comparison: Stochastic gradient against zero order methods for logistic regression:

$$f(\theta) = \frac{1}{n}\sum_{i=1}^n \log\left(1 + \exp(-y_i x_i^\top \theta)\right).$$



Optimality gap for SGD and zero-order plotted vs. iteration count



Optimality gap for SGD and zero-order with iteration count rescaled by dimension $d = 50$

- Performance matches theoretical predictions.
- Similar story across wide range of dimensions d.