Necessary and Sufficient Conditions for using Adaptive, Mirror, and Stochastic Gradient Methods

Daniel Levy, John Duchi

Stanford University

Introduction

- For $\Theta \subset \mathbf{R}^d$ convex, compact set, P distribution on $\mathcal{X}$ and $F : \Theta \times \mathcal{X} \to \mathbf{R}$, stochastic optimization aims to solve:
$$\underset{\theta \in \Theta}{\text{minimize}} \;\; f(\theta) := \mathbf{E}_P \left[F(\theta, X)\right] = \int F(\theta, x) dP(x).$$
- Central problem of statistical learning and estimation (e.g. P the data distribution, Θ the set of classifiers and F the convex loss function).
- Often tackled with stochastic gradient methods because of simplicity and scalability **but** poor convergence rates for many constraints set (e.g. when Θ is an ℓ_1 ball).
- This work provides concrete recommendations for when to use **adaptive**, **mirror** or **stochastic** gradient methods.

Notation and Definitions

d is the dimension, n is the number of samples. For γ a norm, $\mathbf{B}_\gamma(x_0, r) := \{x, \gamma(x - x_0) \le r\}$. For h a distance generating function (dgf) $\mathrm{D}_h(x, y) := h(x) - h(y) - \nabla h(y)^\top (x - y)$. $\mathcal{F}^{\gamma,r} := \left\{F : \mathbf{R}^d \times \mathcal{X} \to \mathbf{R} \mid \text{for all } \theta \in \mathbf{R}^d, g \in \partial_\theta F(\theta, x), \gamma(g) \le r\right\}$. A set Θ is **quadratically convex** (QC) if, $\Theta^2 := \{(\theta_j^2)_{j \le d}, \theta \in \Theta\}$ is convex.

Summary of Results

- When Θ is QC and $\mathbf{B}_\gamma(0, 1)$ is QC then diagonally-rescaled stochastic gradient methods are minimax rate optimal.
- When Θ is QC and $\gamma(g) := \|\beta \odot g\|_p$ for $p \ge 1$, then diagonally-rescaled stochastic gradient methods are minimax rate optimal.
- When Θ is **not** QC, the best linearly-preconditioned gradient methods can be arbitrary suboptimal (up to $\sqrt{d/\log d}$) and non-linear mirror descent are minimax rate optimal.
- For $\Theta = \mathcal{B}_\infty$ and $\gamma(g) = \|\beta \odot g\|_1$, stochastic gradient methods can be $\sqrt{d}$ suboptimal compared to AdaGrad – *see paper.*

Background: Algorithms and Regret Bounds

Algorithms For a sample $X_1^n \overset{\text{iid}}{\sim} P$, for $\alpha > 0$ a stepsize and h_i a dgf, first-order methods iteratively set
$$g_i \in \partial_\theta F(\theta_i, X_i), \qquad \theta_{i+1} := \underset{\theta \in \Theta}{\operatorname{argmin}} \left\{ g_i^\top \theta + \frac{1}{\alpha} \mathrm{D}_h(\theta, \theta_i) \right\}.$$
For various h_i, we obtain familiar algorithms:

- If $h_i(\theta) = \frac{1}{2}\|\theta\|_2^2$ and $\Theta = \mathbf{R}^d$, $\theta_{i+1} = \theta_i - \alpha g_i$, this is the classical **stochastic gradient method**.
- If $h_i(\theta)$ is a fixed, strongly-convex dgf w.r.t. $\|\cdot\|$, this is **mirror descent** [2].
- If $h_i(\theta) = \frac{1}{2}\theta^\top G_t \theta$ with $G_t := \operatorname{diag}\left(\sum_{l \le i} g_l g_l^\top\right)^{1/2}$, this is **AdaGrad** [4].

Regret Bound For $\theta_1, \ldots, \theta_n$ played on functions $\{F(\cdot, x_i)\}_{i \le n}$, the regret w.r.t. θ is $\mathsf{Regret}_n(\theta) := \sum_{i=1}^n [F(\theta_i, x_i) - F(\theta, x_i)]$. When playing θ_1^n as above, the following holds
$$\mathsf{Regret}_n(\theta) \le \frac{\mathrm{D}_h(\theta, \theta_0)}{\alpha} + \frac{\alpha}{2} \sum_{i \le n} \|g_i\|_*^2.$$

Background: Minimax rates

Complexity of problems is measured via **minimax rates** [1]. Let Θ be closed, convex set, $\mathcal{X}$ a sample space, $\mathcal{F}$ a family of functions and $\mathcal{P}$ a family of distributions over $\mathcal{X}$. The minimax stochastic risk is
$$\mathfrak{M}_n^{\mathsf{S}}(\Theta, \mathcal{F}, \mathcal{P}) := \inf_{\widehat{\theta}_n} \sup_{F \in \mathcal{F}} \sup_{P \in \mathcal{P}} \mathbf{E}\left[f_P(\widehat{\theta}_n(X_1^n)) - \inf_{\theta \in \Theta} f_P(\theta) \right].$$
Intuitively, it corresponds to **the best algorithm** given samples X_1^n on the **hardest problem**. Related notion: (average) **minimax regret** where point $\widehat{\theta}_i$ is chosen conditional on x_1^{i-1}:
$$\mathfrak{M}_n^{\mathsf{R}}(\Theta, \mathcal{F}, \mathcal{X}) := \frac{1}{n} \inf_{\widehat{\theta}_{1:n}} \sup_{F \in \mathcal{F}, x_1^n \in \mathcal{X}^n, \theta \in \Theta} \mathsf{Regret}_n(\theta).$$
For a given norm γ, consider $\mathcal{F} = \mathcal{F}^{\gamma,1}$ – the geometries of γ and Θ determine the minimax regret and risk. Given that $\mathfrak{M}_n^{\mathsf{R}}(\Theta, \gamma) \le \mathfrak{M}_n^{\mathsf{S}}(\Theta, \gamma)$ [3], we lower bound the LHS and upper bound the RHS. **When those match, we found the minimax optimal estimator.**

Quadratically Convex Constraint Sets

Let Θ be a QC, orthosymmetric, convex and compact set.

1. If γ is QC, then
$$\mathfrak{M}_n^{\mathsf{R}}(\Theta, \gamma) \asymp \mathfrak{M}_n^{\mathsf{S}}(\Theta, \gamma) \asymp \frac{1}{\sqrt{n}} \sup_{\theta \in \Theta} \gamma^*(\theta).$$
2. If $\gamma(g) = \|\beta \odot g\|_p$ for $p \in [1, 2]$ and $\beta \succ 0$, then
$$\mathfrak{M}_n^{\mathsf{R}}(\Theta, \gamma) \asymp \mathfrak{M}_n^{\mathsf{S}}(\Theta, \gamma) \asymp \frac{1}{\sqrt{n}} \sup_{\theta \in \Theta} \|\theta/\beta\|_2.$$

Moreover, these are attained by diagonal gradient descent. For the lower bound, we find the hardest rectangular sub-problem. The upper bound relies on strong duality which crucially holds because of quadratic convexity. **Gradient methods with a fixed, diagonal pre-conditioner are optimal on such problems.**

Beyond Quadratic Convexity

For $p \in [1, 2]$, we consider $\Theta = \mathcal{B}_p$ and $\gamma = \ell_{p^*}$ for $1/p + 1/p^* = 1$. We have

1. If $1 \le p \le 1 + 1/\log(2d)$, $\mathfrak{M}_n^{\mathsf{S}}(\Theta, \gamma) \asymp \mathfrak{M}_n^{\mathsf{R}}(\Theta, \gamma) \asymp \sqrt{\frac{\log(2d)}{n}}$.
2. If $1 + 1/\log(2d) < p \le 2$, $\mathfrak{M}_n^{\mathsf{S}}(\Theta, \gamma) \asymp \mathfrak{M}_n^{\mathsf{R}}(\Theta, \gamma) \asymp \sqrt{\frac{1}{n(p-1)}}$.

In either case, the upper bound corresponds to (non-linear) mirror descent with dgf $h(\theta) := \frac{1}{2(a-1)}\|\theta\|_a^2$ with, for *(1)* $a = 1 + \frac{1}{\log(2d)}$, for *(2)* $a = p$. We exhibit problems where standard gradient methods achieve their upper bound regret and characterize the suboptimality gap with mirror descent. When p is very close to 2 (i.e. very close to QC), the gap is a constant factor, when $p = 1$, the gap is $\sqrt{d/\log d}$. **In high dimensions, Euclidean gradient methods are arbitrarily suboptimal on this class of problems.**

[1] Alekh Agarwal, Peter L. Bartlett, Pradeep Ravikumar, and Martin J. Wainwright. Information-theoretic lower bounds on the oracle complexity of convex optimization. *IEEE Transactions on Information Theory*, 58(5):3235–3249, 2012.

[2] A. Beck and M. Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31:167–175, 2003.

[3] N. Cesa-Bianchi, A. Conconi, and C. Gentile. On the generalization ability of on-line learning algorithms. In *Advances in Neural Information Processing Systems 14*, pages 359–366, 2002.

[4] John C. Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12:2121–2159, 2011.