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

Daniel Levy, John Duchi
Stanford University

Introduction

For Θ ⊂ R^d convex, compact set, P distribution on X and F : Θ × X → R, stochastic optimization aims to solve:

\[ \minimize f(θ) = \mathbb{E}_P[F(θ, X)] = \int F(θ, x) dP(x). \]

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 (e.g. when Θ is an ℓ1-balls).

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, \( B_d(x_0, r) := \{ x \mid γ(x - x_0) ≤ r \} \). For \( h \) a distance generating function (dgf) \( D_h(x, y) := h(x) - h(y) - \nabla h(y)^\top (x - y) \)

\[ F := \{ F : \mathbb{R}^d \times X \rightarrow \mathbb{R} \mid \text{for all } θ \in \mathbb{R}^d, g \in \partial h(F(θ, x), \gamma(x) \leq r) \}. \]

A set \( Θ \) is quadratically convex (QC) if \( Θ^2 := \{ (θ_0, θ) \in Θ \} \) is convex.

Background: Algorithms and Regret Bounds

Algorithms For a sample \( X^n := P \), for \( α > 0 \) a stepsize and \( h_i \) a dfg, first-order methods iteratively set

\[ g_{i+1} := \arg\min_{g \in \partial h(F(θ_i, X_i))} \left\{ g^\top θ_i + \frac{1}{α} D_h(θ_i, θ_i) \right\}. \]

For various \( h_i \), we obtain familiar algorithms:

- If \( h_i(θ) = \frac{1}{2}∥θ∥^2 \) and \( Θ = \mathbb{R}^d \), \( θ_{i+1} = θ_i - α g_i \), this is the classical stochastic gradient method.
- If \( h_i(θ) \) is a fixed, strongly-convex dfg w.r.t \( ∥·∥\), this is mirror descent [2].
- If \( h_i(θ) = θ^\top G θ \) with \( G_i := \text{diag} \left( ∑_{j \leq i} g_j^\top g_j \right) ^{1/2} \), this is AdaGrad [4].

Regret Bound For \( θ_1, ..., θ_n \) played on functions \( \{ F(θ, x_i) \}_{i \leq n} \), the regret w.r.t. \( θ \) is

\[ \text{Regret}_n(θ) := ∑_{i \leq n} F(θ_i, x_i) - F(θ, x_i). \]

When playing \( θ_i^∗ \) as above, the following holds

\[ \text{Regret}_n(θ) ≤ D_h(θ_i, θ_i^∗) + α \frac{α}{2} ∑_{i \leq n} ∥g_i∥^2. \]

Background: Minimax rates

Complexity of problems is measured via minimax rates [1]. Let Θ be closed, convex set, \( X \) a sample space, \( F \) a family of functions and \( P \) a family of distributions over \( X \). The minimax stochastic risk is

\[ M^*_{\mathbb{E}}(Θ, F, P) := \inf_{θ \in Θ} \sup_{F \in F, P \in P} \mathbb{E} \left[ f_p(θ_p(X^n_i)) - f_p(θ) \right]. \]

Intuitively, it corresponds to the best algorithm given samples \( X^n \). On the hardest problem, related notion: (average) minimax regret

\[ \text{Regret}^*_n(θ) := \inf_{θ \in Θ} \sup_{F \in F, P \in P, θ_p \in Θ} \text{Regret}_n(θ). \]

For a given norm \( γ \), consider \( F = F^{θ^*} \), the geometry of \( γ \) and \( θ^* \) determine the minimax regret and risk. Given that \( M^*_{\mathbb{E}}(Θ, γ) ≤ M^*_{\mathbb{E}}(Θ, γ) \) [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.

- If \( γ \) is QC, then

\[ M^*_{\mathbb{E}}(Θ, γ) ≤ M^*_{\mathbb{E}}(Θ, γ) ≤ \frac{1}{\sqrt{n}} \sup_{θ \in Θ} γ^∗(θ). \]

- If \( γ(θ) = ∥θ∥_p \) for \( p \in [1, 2] \) and \( p > 0 \), then

\[ M^*_{\mathbb{E}}(Θ, γ) ≤ M^*_{\mathbb{E}}(Θ, γ) ≤ \frac{1}{\sqrt{n}} \sup_{θ \in Θ} ∥θ∥_p. \]

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 \( Θ = B_p \) and \( γ = ℓ_p \). for \( 1/p + 1/p^* = 1 \). We have

- If \( 1 ≤ p ≤ 1 + 1/\log(2d) \), \( M^*_{\mathbb{E}}(Θ, γ) ≤ M^*_{\mathbb{E}}(Θ, γ) ≤ \frac{n^{1/p}}{m^{1/p}} \)

- If \( 1 + 1/\log(2d) < p ≤ 2 \), \( M^*_{\mathbb{E}}(Θ, γ) ≤ M^*_{\mathbb{E}}(Θ, γ) ≥ \frac{n^{1/p}}{m^{1/p}} \).

In either case, the upper bound corresponds to (non-linear) mirror descent with \( d(θ) := \frac{1}{∥θ∥_p} \) with, for (1) \( a = 1 + \frac{1}{1/p} \), 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 \( d/\log d \). In high dimensions, Euclidean gradient methods are arbitrarily suboptimal on this class of problems.