A Characterization of Stochastic Mirror Descent Algorithms and Their Convergence Properties

Navid Azizan and Babak Hassibi

Motivation
- **Stochastic Mirror Descent (SMD)** is a general family of optimization algorithms
- **Stochastic Gradient Descent (SGD)** is a special case of SMD
- Other examples include exponential weights algorithm, p-norms algorithm, etc.
- SMD algorithms have become increasingly popular in optimization, machine learning, signal processing, control, etc.

Problem Setup
- Data: \( \{(x_i, y_i) : i = 1, \ldots, n\} \)
  where \( x_i \in \mathbb{R}^d \) and \( y_i \in \mathbb{R} \)
- Model: \( y_i = f(x_i, w) + \eta_i, \quad i = 1, \ldots, n \)
  \( f(\cdot, \cdot) \) is a given function that represents the model class
  \( w \in \mathbb{R}^p \) is an unknown weight vector (parameter)
  \( \eta_i \) is the noise, which represents measurement error, modeling error, etc.
- Loss Function: \( l(\cdot) \) is a nonnegative differentiable loss function with \( l(0) = 0 \)
  \( L(w) = \sum_{i=1}^n l_i(w) \)
- SGD: \( w_t = w_{t-1} - \eta \nabla L(w_{t-1}) \)

Minimax Optimality of SGD
- Consider a linear model \( f(\cdot, w) = x^T w, i.e., y_i = x_i^T w + \nu_i \)
  and the square loss \( L(w) = \frac{1}{2} \|x - x w \|^2 \)
- In this case, SGD is \( w_t = w_{t-1} - \eta \nabla L(w_{t-1}) \)

Theorem (Hassibi et al., NIPS ’09): For any initialization \( w_0 \), any sufficiently small step size \( \eta \), i.e., \( 0 < \eta \leq \min \frac{1}{\|x^T x\|_2} \), and any number of steps \( T \geq 1 \), the SGD iterates \( \{w_t\} \) are the optimal solution to the following minimization problem

\[
\min_{\{w_t \}} \max_{\{v_t \}} \frac{1}{T} \sum_{t=1}^T \|w_t - v_t\|^2 + \eta \sum_{t=1}^T \|x_t - x \|^2 \]

and the optimal value is \( 1 \).

- The ratio is the \( \mathbb{F}_{\text{TV}} \) norm of the transfer operator that maps the unknown disturbances to the estimation errors
- Interpretations: Robustness and Conservatism

Proof: The Conservation Law of SGD
Define “innovations” and “predicted error” as \( e_{i} = y_i - x_i^T w_{i-1} \)
and \( p_{i} = x_i^T w - x_i^T w_{i-1} \)

Conservation of Uncertainty

For each step of SGD:

\[
\|w - w_{i-1}\|^2 + \eta p_i^2 = \|w - w_{i-1}\|^2 + \eta (1 - \eta \|x_i\|^2) p_i^2 + \eta p_i^2, \quad \forall i \geq 1.
\]

Lemma. For any noise values \( \{\eta_i\} \), any true parameter \( w \), and any step-size sequence \( \{\eta_i\} \), the following relation holds for the SGD iterates \( \{w_i\} \)

Implications for Overparameterized Models
- Set of solutions: \( W = \{w_i \mid y_i = w_i x_i, \quad i = 1, \ldots, n\} \)
- Convergence and Implicit Regularization:
  \[
  \text{For } \eta < \min \frac{1}{\|x_i\|^2}, \quad \text{the SGD iterates converge to a solution } w_\infty \in W. \quad \text{Further}
  w_\infty = \arg \min \{w \mid \|w - w_0\| \}
  \]
  \[
  \text{In particular, if initialized at zero, SGD converges to the minimum } \ell_2 \text{ norm solution}
  w_\infty = \arg \min \{w \mid \|w\| \}
  \]
  This is called implicit regularization.

What if we want a different regularizer?

Stochastic Mirror Descent (SMD)
- A general family of optimization algorithms that includes stochastic gradient descent
- For any strictly convex and differentiable potential \( \psi \), the SMD update rule is
  \[
  w_i = \arg \min \psi(w) - \psi(w_{i-1}) + \frac{\eta}{2} \|w - w_{i-1}\|^2
  \]
  where \( D_\psi(w_{i-1}) = \nabla \psi(w_{i-1}) - \nabla \psi(w_{i-1})^T (w - w_{i-1}) \) is the Bregman divergence w.r.t. \( \psi \)
- Equivalently, the SMD update can be expressed as
  \[
  \psi(w_i) = \psi(w_{i-1}) - \nabla \psi(w_{i-1})^T (w - w_{i-1})
  \]
- For SGD \( \psi(w) = \frac{1}{2} \|w\|^2 \)

Minimax Optimality of SMD
- Consider any (nonlinear) model \( f(\cdot, w) \), any differentiable loss \( l(\cdot) \)
  with property \( l(0) = l(0) = 0 \) and any initialization \( w_0 \). For sufficiently small sequence of step sizes \( \{\eta_i\}, i.e., one for which \( \psi(w) - \psi(w_{i-1}) \) is convex for all i, and for any number of steps \( T \geq 1 \), the SMD iterates \( \{w_t\} \) are the optimal solution to the following minimization problem

\[
\min_{\{w_t \}} \max_{\{v_t \}} \frac{1}{T} \sum_{t=1}^T \|w_t - v_t\|^2 + \eta \sum_{t=1}^T \|x_t - x \|^2
\]

and the optimal value is \( 1 \).

- Generalizes several results, e.g., SGD/squareless/linear model [Hassibi et al. 2010] and \( \ell_1 \)-norm/squareless/linear model [Simsekli et al. 2012]
- Proof by the conservation law of SGD:

Lemma. For any model \( f(\cdot, \cdot) \), any differentiable loss \( l(\cdot) \), any parameter \( w \) and noise values \( \{\nu_i\} \) that satisfy \( y_i = f(x_i, w) + \nu_i \) for \( i = 1, \ldots, n \), and any step-size sequence \( \{\eta_i\} \), the following relation holds for the SMD iterates

\[
D_\psi(w_{i-1}) + \eta \nu_i = D_\psi(w_i) + E_\psi(w_{i-1}) + \eta D_\psi(w_{i-1}) + \eta L(w_{i-1}),
\]

for all \( i \geq 1 \), where \( E_\psi(w_{i-1}) = D_\psi(w_{i-1}) - D_\psi(w_{i-1}) + \eta D_\psi(w_{i-1}) + \eta L(w_{i-1}) \).

Implicit Regularization of SMD
- Proposition. If \( l \) is differentiable and convex and has a unique root at 0, \( \psi \) is strictly convex, and positive sequence \( \{\eta_i\} \) such that \( \psi - \psi(w_i) \) is convex for all \( i \), then for any initialization \( w_0 \), the SMD iterates converge to

\[
\psi(w_\infty) = \arg \min \{w \mid \psi(w) \}
\]

In particular, if we initialize SMD with \( w_0 = \arg \min \psi(w) \), then \( w_\infty \) is the minimization solution. This is another implicit regularization.

One can choose the potential function of SMD for any desired regularization.

Example: Compressed Sensing via SMD
- Recovering a sparse signal

\[
\min \|w\|_1 \quad \text{s.t.} \quad x = \Phi w + \nu
\]

where \( \Phi \) is a random Gaussian sensing matrix and \( \nu \) is i.i.d. Gaussian noise.

SMD w. \( \psi(w) = \|w\|_1 \) recovers the sparse signal!

Stochastic Convergence in Underparameterized models
- Under-parameterized (online streaming) linear regression
- Vanishing step size
- Classical result:

Proposition. Consider \( y_i = x_i^T w + \nu_i, \quad i \geq 1 \), where \( E[\nu_i] = 0, \quad E[\nu_i \nu_j] = \sigma^2 \delta_{ij} \), and the \( x_i \) are persistently exciting. For any step size sequence \( \{\eta_i\} \) such that \( \sum_{i=1}^n \eta_i = \infty, \quad \sum_{i=1}^n \eta_i^2 < \infty \), the SMD iterates with respect to any strongly convex potential \( \psi(\cdot) \), converge to \( w \) in the mean-square sense.

- Direct and elementary proof using the conservation law of SMD
- Avoids ergodic averaging or appealing to stochastic differential equations

New Experimental Results
- SMD with different potential functions ran on MNIST
- The problem is non-linear

- 6 initial points x 4 different mirrors x 24 points on the manifold
- Bregman divergences between the final and initial points, in 4 different norms

SMD converges to the point with smallest Bregman divergence from the initial point