A Characterization of Stochastic Mirror Descent Algorithms and Their Convergence Properties

Navid Azizan and Babak Hassibi

Caltech

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: $\{(x_i, y_i) : i = 1, \dots, n\}$
where $x_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$

Model: $y_i = f(x_i, w) + v_i, \quad i = 1, \dots, n$
where $f(\cdot,\cdot)$ is a given function that represents the model class
$w \in \mathbb{R}^p$ is an unknown weight vector (parameter)
v_i is the noise, which represents measurement error, modeling error, etc.

Loss Function: $l(\cdot)$ is a nonnegative differentiable loss function with $l(0) = 0$

$$L_i(w) = l(y_i - f(x_i, w)) \qquad L(w) = \sum_{i=1}^n L_i(w)$$

SGD: $w_i = w_{i-1} - \eta_i \nabla L_i(w_{i-1})$

Minimax Optimality of SGD

Consider a linear model $f(x_i, w) = x_i^T w$, i.e., $y_i = x_i^T w + v_i$,
and the square loss $L_i(w) = \frac{1}{2}(y_i - x_i^T w)^2$
In this case, SGD is $w_i = w_{i-1} + \eta(y_i - x_i^T w_{i-1}) x_i$

Theorem (Hassibi et al, NIPS '93). *For any initialization w_0, any sufficiently small step size η, i.e., $0 < \eta \le \min_i \frac{1}{\|x_i\|^2}$, and any number of steps $T \ge 1$, the SGD iterates $\{w_i\}$ are the optimal solution to the following minimization problem*

$$\min_{\{w_i\}} \max_{w, \{v_i\}} \frac{\|w - w_T\|^2 + \eta \sum_{i=1}^T (x_i^T w - x_i^T w_{i-1})^2}{\|w - w_0\|^2 + \eta \sum_{i=1}^T v_i^2},$$

and the optimal value is 1.

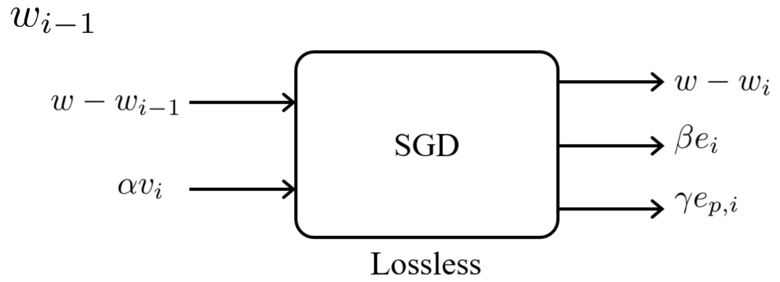
- The ratio is the H^∞ norm of the transfer operator that maps the unknown disturbances to the estimation errors
- Interpretations: Robustness and Conservatism

Proof: The Conservation Law of SGD

Define "innovations" and "predicted error" as
$e_i := y_i - x_i^T w_{i-1} \quad$ and $\quad e_{p,i} := x_i^T w - x_i^T w_{i-1}$

Conservation of Uncertainty

For each step of SGD:



$\alpha = \gamma = \sqrt{\eta_i}$ and $\beta = \sqrt{\eta_i(1 - \eta_i\|x_i\|^2)}$

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

$$\|w - w_{i-1}\|^2 + \eta_i v_i^2 = \|w - w_i\|^2 + \eta_i \left(1 - \eta_i \|x_i\|^2\right) e_i^2 + \eta_i e_{p,i}^2, \quad \forall i \ge 1.$$

Implications for Overparameterized Models

Set of solutions: $\mathcal{W} = \{w \mid y_i = x_i^T w, \ i = 1, \dots, n\}$

Convergence and Implicit Regularization:

For $\eta < \min_i \frac{2}{\|x_i\|^2}$, the SGD iterates converge to a solution $w_\infty \in \mathcal{W}$. Further

$$w_\infty = \arg\min_{w \in \mathcal{W}} \|w - w_0\|$$

In particular, if initialized at zero, SGD converges to the minimum l_2 norm solution $w_\infty = \arg\min_{w \in \mathcal{W}} \|w\|$

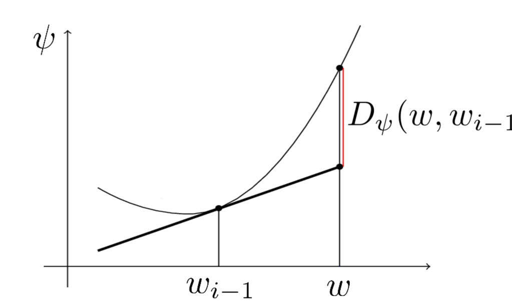
This is called implicit regularization

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 ψ, the SMD update rule is $w_i = \arg\min_w \eta w^T \nabla L_i(w_{i-1}) + D_\psi(w, w_{i-1})$

  where $D_\psi(w, w_{i-1}) = \psi(w) - \psi(w_{i-1}) - \nabla\psi(w_{i-1})^T (w - w_{i-1})$ is the Bregman divergence w.r.t. ψ



- Equivalently, the SMD update can be expressed as

$$\nabla\psi(w_i) = \nabla\psi(w_{i-1}) - \eta_i \nabla L_i(w_{i-1})$$

- For SGD $\psi(w) = \frac{1}{2}\|w\|^2$

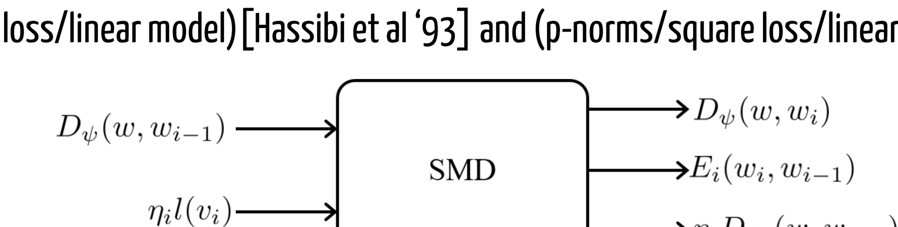
Minimax Optimality of SMD

Theorem. *Consider any (nonlinear) model f, any differentiable loss l 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 $\psi(w) - \eta_i L_i(w)$ is convex for all i, and for any number of steps $T \ge 1$, the SMD iterates $\{w_i\}$, w.r.t. any strictly convex potential ψ, are the optimal solution to the following minimization problem*

$$\min_{\{w_i\}} \max_{w, \{v_i\}} \frac{D_\psi(w, w_T) + \sum_{i=1}^T \eta_i D_{L_i}(w, w_{i-1})}{D_\psi(w, w_0) + \sum_{i=1}^T \eta_i l(v_i)},$$

and the optimal value is 1.

- Generalizes several results, e.g. (SGD/square loss/linear model) [Hassibi et al '93] and (p-norms/square loss/linear model) [Kivinen et al '06]
- Proof by the conservation law of SMD:



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

$$D_\psi(w, w_{i-1}) + \eta_i l(v_i) = D_\psi(w, w_i) + E_i(w_i, w_{i-1}) + \eta_i D_{L_i}(w, w_{i-1}),$$

for all $i \ge 1$, where $E_i(w_i, w_{i-1}) := D_\psi(w_i, w_{i-1}) - \eta_i D_{L_i}(w_i, w_{i-1}) + \eta_i L_i(w_i)$.

Implicit Regularization of SMD

Proposition. *If l is differentiable and convex and has a unique root at 0, ψ is strictly convex, and positive sequence $\{\eta_i\}$ is such that $\psi - \eta_i L_i$ is convex for all i, then for any initialization w_0, the SMD iterates converge to*

$$w_\infty = \arg\min_{w \in \mathcal{W}} D_\psi(w, w_0).$$

In particular, if we initialize SMD with $w_0 = \arg\min_{w \in \mathbb{R}^m} \psi(w)$, it converges to $w_\infty = \arg\min_{w \in \mathcal{W}} \psi(w)$
i.e., the minimum-potential solution. This is another implicit regularization.

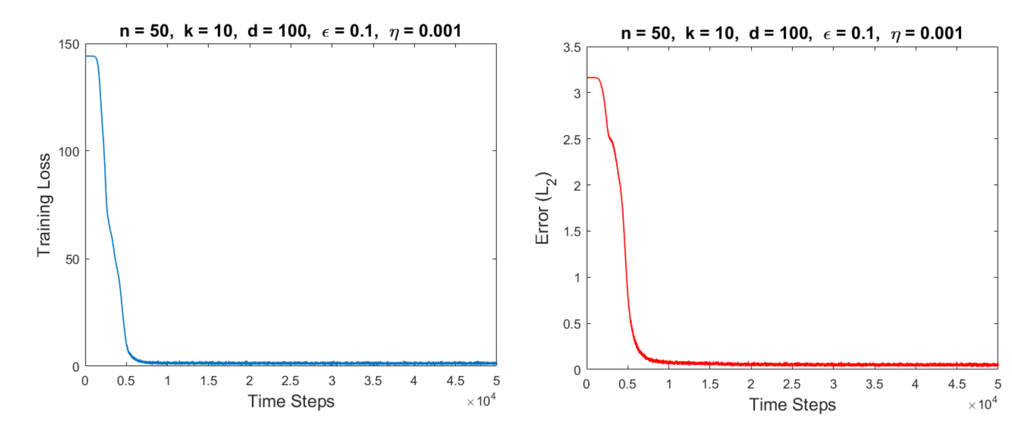
One can choose the potential function of SMD for any desired regularization

Example: Compressed Sensing via SMD

recovering a sparse signal

$$\begin{aligned} \min_w \quad & \|w\|_1 \\ \text{s.t.} \quad & x_i^T w = y_i, \ i = 1, \dots, n. \end{aligned}$$

$\psi(w) = \|w\|_1$ is not differentiable
but we can use $\psi(w) = \|w\|_{1+\epsilon}^{1+\epsilon}$:



SMD w. $\psi(w) = \|w\|_{1+\epsilon}$ recovers the sparse solution!

Stochastic Convergence in Underparameterized models

- Under-parameterized (online streaming) linear regression
- Vanishing step size
- Classical result:

Proposition. *Consider $y_i = x_i^T w + v_i, i \ge 1$, where $\mathbb{E}[v_i] = 0$, $\mathbb{E}[v_i v_j] = \sigma^2 \delta_{ij}$, and the x_i are persistently exciting. For any step size sequence $\{\eta_i\}$ such that $\sum_{i=1}^\infty \eta_i = \infty, \sum_{i=1}^\infty \eta_i^2 < \infty$, 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**

	Final 2 o.b. 1-nor	Final 4 o.b. 1-no	Final 5 o.b. 1-no	Final 6 o.b. 1-n	Final 7 o.b. 1-n	Final 8 o.b. 1-n	Final 2 o.	Final 4 o.	Final 5 o.	Final 6 o.	Final 7 o.	Final 8 o.
Initial 2 BD w/t 1-norm	3.18	20301.98	20302.34	20296.08	20337.84	20341.86	21051	20496	20493	20530	20489	20485
Initial 4 BD w/t 1-norm	20450.91	2.79	20286.68	20261.06	20319.39	20361.36	20593	929.7	20487	20490	20467	20504
Initial 5 BD w/t 1-norm	20456.87	20288.43	2.8196	20255.06	20301.18	20341.49	20598	20477	933.9	20488	20466	20491
Initial 6 BD w/t 1-norm	20486.22	20292.17	20281.79	2.9139	20295.74	20369.33	20633	20479	20473	927.2	20451	20510
Initial 7 BD w/t 1-norm	20502.83	20299.8	20276.66	20256.43	2.7454	20341.65	20647	20493	20468	20483	942.1	20477
Initial 8 BD w/t 1-norm	20467.21	20308.67	20314.69	20287.48	20304.85	2.5929	20619	20494	20512	20523	20464	924
Initial 2 BD w/t 2-norm	316.15	931.63	918.5897	931.03	930.98	901.2859	64.61	670.1	667.2	670.4	669.8	668.8
Initial 4 BD w/t 2-norm	956.36	295.92	920.7616	929.2809	930.5393	902.885	675.1	59.2924	668.1	669.64	669.9	671.7
Initial 5 BD w/t 2-norm	955.14	930.3507	288.3069	929.3628	927.5	902.8	672.8	667.8	58.82	667.9	667.7	670.8
Initial 6 BD w/t 2-norm	956.34	931.62	921.12	295.7218	927.47	902.92	675.3	669.8	667.8	57.89	667.1	671.3
Initial 7 BD w/t 2-norm	958.26	931.26	920.64	927.04	294.57	901.24	676.7	670	668.4	667.3	60.43	670
Initial 8 BD w/t 2-norm	953.92	933.37	921.33	931.38	928.11	263.2133	673.9	672.2	669.4	671.3	669.2	58.59
Initial 2 BD w/t 3-norm	1852.6	1582.03	1480.255	1567.559	1551.939	1381.913	39.14	68.77	65.94	65.93	68.88	67.87
Initial 4 BD w/t 3-norm	1884.36	1548.95	1482.5909	1566.449	1551.91	1379.24	67.34	39.65	66.23	65.66	68.95	67.83
Initial 5 BD w/t 3-norm	1885.68	1582.85	1453.5978	1568.656	1550.95	1382.347	66.89	68.25	38.61	65.68	68.63	68.36
Initial 6 BD w/t 3-norm	1883.29	1582.77	1483.83	1536.133	1550.46	1379.26	67.43	68.8	66.64	37.05	68.15	67.76
Initial 7 BD w/t 3-norm	1884.19	1580.58	1482.96	1565.3	1520.26	1380.117	67.42	68.38	66.72	64.84	40.61	67.97
Initial 8 BD w/t 3-norm	1881.79	1581.19	1481.81	1566.21	1549.15	1345.4	66.78	68.92	66.43	65.56	68.86	38.74
Initial 2 BD w/t 10-norm	1.81E+09	1.31E+08	1.02E+08	1.27E+08	1.08E+08	1.30E+08	818.6	869.5	801.9	768.1	935.2	811.3
Initial 4 BD w/t 10-norm	1.81E+09	1.31E+08	1.02E+08	1.27E+08	1.08E+08	1.30E+08	819.1	868.7	802	767.9	935.2	811
Initial 5 BD w/t 10-norm	1.81E+09	1.31E+08	1.02E+08	1.27E+08	1.08E+08	1.30E+08	819.1	869.5	801.7	768.1	935.2	811.3
Initial 6 BD w/t 10-norm	1.81E+09	1.31E+08	1.02E+08	1.27E+08	1.08E+08	1.30E+08	819	869.4	802.1	767.4	935.1	811
Initial 7 BD w/t 10-norm	1.81E+09	1.31E+08	1.02E+08	1.27E+08	1.08E+08	1.30E+08	819	869.3	802.1	767.8	934.6	811.1
Initial 8 BD w/t 10-norm	1.81E+09	1.31E+08	1.02E+08	1.27E+08	1.08E+08	1.30E+08	818.8	869.2	801.9	767.6	935.1	810.4

	Final 2 o.	Final 4 o.	Final 5 o.	Final 6 o.	Final 7 o.	Final 8 o.	Final 2 o.	Final 4 o.	Final 5 o.	Final 6 o.	Final 7 o.	Final 8 o.
Initial 2 BD w/t 1-norm	10464	22680	23306	22675	22759	22615	2E+05	2E+05	2E+05	2E+05	2E+05	2E+05
Initial 4 BD w/t 1-norm	22693	10340	23300	22623	22720	22614	2E+05	2E+05	2E+05	2E+05	2E+05	2E+05
Initial 5 BD w/t 1-norm	22694	22653	11637	22608	22748	22613	2E+05	2E+05	2E+05	2E+05	2E+05	2E+05
Initial 6 BD w/t 1-norm	22726	22652	23288	10337	22701	22634	2E+05	2E+05	2E+05	2E+05	2E+05	2E+05
Initial 7 BD w/t 1-norm	22746	22675	23300	22627	10501	22622	2E+05	2E+05	2E+05	2E+05	2E+05	2E+05
Initial 8 BD w/t 1-norm	22716	22682	23334	22654	22718	10343	2E+05	2E+05	2E+05	2E+05	2E+05	2E+05
Initial 2 BD w/t 2-norm	268.06	701.9	714.2	708.248	702.9	701	30279	30768	30311.7	29513	30092	31508
Initial 4 BD w/t 2-norm	703	263.5	715.5	706.8	702.5	704.8	30252	30788	30310	29508	30097	31513
Initial 5 BD w/t 2-norm	701.086	699.7	296.1	705.5	702.1	702.8	30252	30761	30331	29514	30092	31503
Initial 6 BD w/t 2-norm	702.9	701.6	715.2	272.7	700.1	704.6	30250	30768	30310	29525	30096	31520
Initial 7 BD w/t 2-norm	704.3	702	716	705	268.2	703.6	30256	30769	30310	29510	30122	31514
Initial 8 BD w/t 2-norm	702.3	705.1	717.3	709.1	702.3	265.9	30254	30769	30318	29505	30097	31528
Initial 2 BD w/t 3-norm	7.6	32.09	31.98	32.75	32.05	32.01	4316	4404	4321	4188	4211	4511
Initial 4 BD w/t 3-norm	32.23	7.399	32.28	32.72	32.54	32.95	4314	4407	4323	4187	4214	4513
Initial 5 BD w/t 3-norm	31.57	31.72	8.302	32.38	32.23	32.42	4313	4403	4324	4188	4211	4511
Initial 6 BD w/t 3-norm	32.52	32.44	32.6	7.994	31.86	32.74	4314	4404	4323	4190	4213	4513
Initial 7 BD w/t 3-norm	32.19	32.23	32.62	32.1	7.577	32.64	4314	4405	4324	4188	4214	4512
Initial 8 BD w/t 3-norm	32.13	33.09	32.87	33.1	32.68	7.59	4314	4405	4323	4188	4213	4513
Initial 2 BD w/t 10-norm	0.3654	0.728	0.654	0.8094	0.718	0.726	0.02	0.36	0.297	0.373	0.333	0.327
Initial 4 BD w/t 10-norm	0.765	0.355	0.7523	0.8	0.794	0.726	0.456	0.02	0.413	0.398	0.441	0.369
Initial 5 BD w/t 10-norm	0.699	0.787	0.4	0.777	0.869	0.746	0.368	0.43	0.02	0.339	0.515	0.347
Initial 6 BD w/t 10-norm	0.76	0.749	0.78	0.431	0.783	0.746	0.458	0.409	0.439	0.018	0.437	0.392
Initial 7 BD w/t 10-norm	0.694	0.761	0.802	0.768	0.38	0.779	0.385	0.43	0.469	0.378	0.016	0.423
Initial 8 BD w/t 10-norm	0.693	0.733	0.756	0.739	0.819	0.38	0.4	0.403	0.426	0.361	0.479	0.021

6 initial points x 4 different mirrors = 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