Dual Averaging Method for Regularized Stochastic Learning and Online Optimization

Lin Xiao
Microsoft Research

NIPS 2009 / NeurIPS 2019

Outline

- background and motivation
- main results of the paper
- further developments in recent years
- final reflections

Stochastic gradient descent (SGD)

stochastic optimization:

$$\underset{w}{\text{minimize}} \quad \mathbf{E}_z f(w, z)$$

empirical risk minimization:

$$\underset{w}{\text{minimize}} \quad \frac{1}{n}\sum_{i=1}^{n} f(w, z_i)$$

SGD: for $t = 0, 1, 2, \ldots$

$$w_{t+1} = w_t - \alpha_t \nabla f(w_t, z_t)$$

- goes back to seminal work of Robbins & Monro (1951)
- many variations, workhorses in machine learning
- 2018 Test of Time award: Bottou & Bousquet (2008)

Stochastic gradient descent (SGD)

stochastic optimization:

$$\underset{w}{\text{minimize}} \quad \mathbf{E}_z f(w, z)$$

empirical risk minimization:

$$\underset{w}{\text{minimize}} \quad \frac{1}{n}\sum_{i=1}^{n} f(w, z_i)$$

SGD: for $t = 0, 1, 2, \ldots$

$$w_{t+1} = w_t - \alpha_t \nabla f(w_t, z_t)$$

- goes back to seminal work of Robbins & Monro (1951)
- many variations, workhorses in machine learning
- 2018 Test of Time award: Bottou & Bousquet (2008)

basic convergence theory:

- $O(1/\sqrt{t})$ rate if $f(\cdot, z)$ convex and $\alpha_t \sim 1/\sqrt{t}$
- $O(1/t)$ rate if $f(\cdot, z)$ strongly convex and $\alpha_t \sim 1/t$

Online convex optimization

input: a convex set S
for $t = 1, 2, 3, \ldots$
  predict a vector $w_t \in \mathcal{S}$
  receive a convex loss function $f_t(\cdot)$
  suffer loss $f_t(w_t)$

regret:

$$R_T = \sum_{t=1}^{T} f_t(w_t) - \min_{w \in \mathcal{S}} \sum_{t=1}^{T} f_t(w)$$

Online convex optimization

input: a convex set S
for $t = 1, 2, 3, \ldots$
  predict a vector $w_t \in \mathcal{S}$
  receive a convex loss function $f_t(\cdot)$
  suffer loss $f_t(w_t)$

regret:

$$R_T = \sum_{t=1}^{T} f_t(w_t) - \min_{w \in \mathcal{S}} \sum_{t=1}^{T} f_t(w)$$

online gradient descent (Zinkevich 2003)

$$w_{t+1} = \mathrm{P}_{\mathcal{S}}\left(w_t - \alpha_t \nabla f_t(w_t)\right)$$

- $R_t = O(\sqrt{T})$ for convex, $O(\ln(T))$ for strongly convex losses
- online to batch conversion: $\frac{1}{T}\sum_{t=1}^{T} w_t$ has $\frac{R_T}{T}$ rate

see surveys by Hazan (2011,2019) and Shalev-Shwartz (2012)

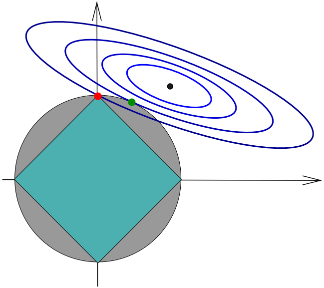
Compressed sensing / sparse optimization

Lasso (Tibshirani 1996):

$$\underset{w}{\text{minimize}} \quad \frac{1}{2}\|Xw - y\|_2^2$$
$$\text{subject to} \quad \|w\|_1 \leq \delta$$

ℓ_1-regularized least-squares:

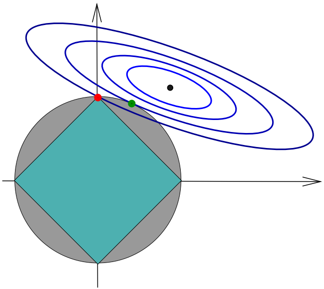
$$\underset{w}{\text{minimize}} \quad \frac{1}{2}\|Xw - y\|_2^2 + \lambda\|w\|_1$$



Compressed sensing / sparse optimization

Lasso (Tibshirani 1996):

$$\begin{array}{ll} \underset{w}{\text{minimize}} & \frac{1}{2}\|Xw - y\|_2^2 \\ \text{subject to} & \|w\|_1 \le \delta \end{array}$$



ℓ_1-regularized least-squares:

$$\underset{w}{\text{minimize}} \quad \frac{1}{2}\|Xw - y\|_2^2 + \lambda\|w\|_1$$

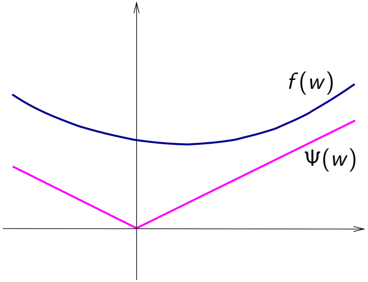
- compressed sensing theory (Donoho, Candès, Tao, ...2004$\sim$)
- generalizations: low-rank matrix completion, nuclear-norm, ...
- algorithms:
  - interior-point methods
  - greedy algorithms: (orthogonal) matching pursuit, LARS, ...
  - proximal gradient methods: ISTA, FISTA, ...

Proximal gradient method

- composite convex optimization

$$\underset{w}{\text{minimize}} \quad f(w) + \Psi(w)$$

  - f convex and smooth
  - Ψ convex and simple

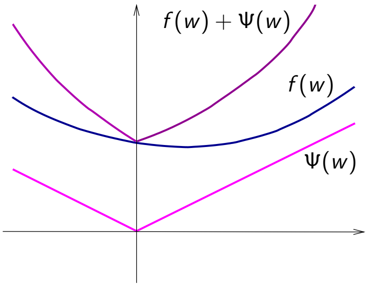


Proximal gradient method

- composite convex optimization

$$\underset{w}{\text{minimize}} \quad f(w) + \Psi(w)$$

  - f convex and smooth
  - Ψ convex and simple



Proximal gradient method

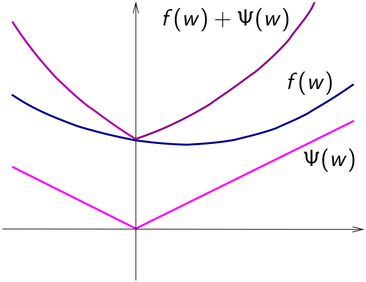
- composite convex optimization

$$\underset{w}{\text{minimize}} \quad f(w) + \Psi(w)$$

  - f convex and smooth
  - Ψ convex and simple

- **subgradient method**

$$w_{t+1} = w_t - \alpha_t \big(\nabla f(w_t) + \xi_t\big)$$

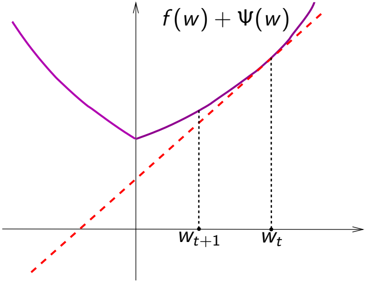


Proximal gradient method

- composite convex optimization

$$\underset{w}{\text{minimize}} \quad f(w) + \Psi(w)$$

  - f convex and smooth
  - Ψ convex and simple



- **subgradient method**

$$w_{t+1} = w_t - \alpha_t \big(\nabla f(w_t) + \xi_t\big)$$

equivalently

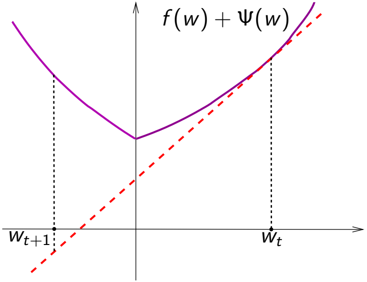
$$w_{t+1} = \arg\min_{w} \left\{ f(w_t) + \langle \nabla f(w_t) + \xi_t, w - w_t \rangle + \frac{1}{2\alpha_t} \|w - w_t\|^2 \right\}$$

Proximal gradient method

- composite convex optimization

$$\underset{w}{\text{minimize}} \quad f(w) + \Psi(w)$$

  - f convex and smooth
  - Ψ convex and simple



- **subgradient method**

$$w_{t+1} = w_t - \alpha_t \big(\nabla f(w_t) + \xi_t\big)$$

equivalently

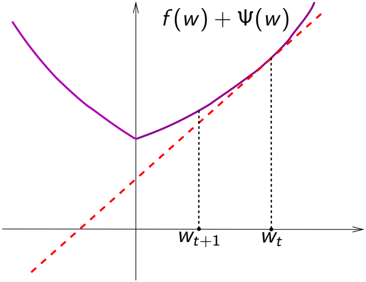
$$w_{t+1} = \arg\min_{w} \left\{ f(w_t) + \langle \nabla f(w_t) + \xi_t, w - w_t \rangle + \frac{1}{2\alpha_t} \|w - w_t\|^2 \right\}$$

Proximal gradient method

- composite convex optimization

$$\underset{w}{\text{minimize}} \quad f(w) + \Psi(w)$$

  - f convex and smooth
  - Ψ convex and simple



- **subgradient method** $\alpha_t \sim 1/\sqrt{t}$

$$w_{t+1} = w_t - \alpha_t\big(\nabla f(w_t) + \xi_t\big)$$

equivalently

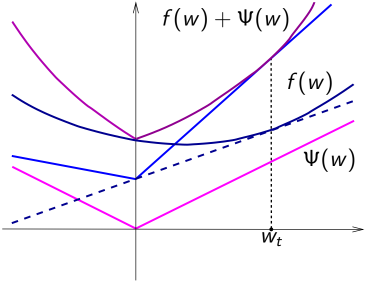
$$w_{t+1} = \arg\min_{w} \left\{ f(w_t) + \langle \nabla f(w_t) + \xi_t, w - w_t \rangle + \frac{1}{2\alpha_t}\|w - w_t\|^2 \right\}$$

Proximal gradient method

- composite convex optimization

$$\underset{w}{\text{minimize}} \quad f(w) + \Psi(w)$$

  - f convex and smooth
  - Ψ convex and simple



- **subgradient method** $\alpha_t \sim 1/\sqrt{t}$

$$w_{t+1} = w_t - \alpha_t \left(\nabla f(w_t) + \xi_t\right)$$

equivalently

$$w_{t+1} = \arg\min_w \left\{ f(w_t) + \langle \nabla f(w_t) + \xi_t, w - w_t \rangle + \frac{1}{2\alpha_t} \|w - w_t\|^2 \right\}$$

- **proximal gradient method**

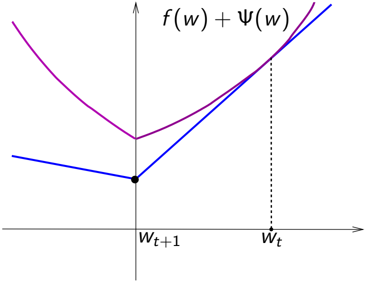
$$w_{t+1} = \arg\min_w \left\{ f(w_t) + \langle \nabla f(w_t), w - w_t \rangle + \Psi(w) + \frac{1}{2\alpha} \|w - w_t\|^2 \right\}$$

Proximal gradient method

- composite convex optimization

$$\underset{w}{\text{minimize}} \quad f(w) + \Psi(w)$$

  - f convex and smooth
  - Ψ convex and simple



- **subgradient method** $\alpha_t \sim 1/\sqrt{t}$

$$w_{t+1} = w_t - \alpha_t \big(\nabla f(w_t) + \xi_t \big)$$

equivalently

$$w_{t+1} = \arg\min_{w} \left\{ f(w_t) + \langle \nabla f(w_t) + \xi_t, w - w_t \rangle + \frac{1}{2\alpha_t} \|w - w_t\|^2 \right\}$$

- **proximal gradient method** (constant α, faster $O(1/t)$ convergence)

$$w_{t+1} = \arg\min_{w} \left\{ f(w_t) + \langle \nabla f(w_t), w - w_t \rangle + \Psi(w) + \frac{1}{2\alpha} \|w - w_t\|^2 \right\}$$

Proximal gradient method

- equivalent form: **forward-backward splitting**

$$
\begin{aligned}
w_{t+\frac{1}{2}} &= w_t - \alpha \nabla f(w_t) \\
w_{t+1} &= \arg\min_{w} \left\{ \alpha \Psi(w) + \frac{1}{2} \left\| w - w_{t+\frac{1}{2}} \right\|_2^2 \right\}
\end{aligned}
$$

or in compact form: $w_{t+1} = \mathbf{prox}_{\alpha\Psi}\big(w_t - \alpha \nabla f(w_t)\big)$

Proximal gradient method

- equivalent form: **forward-backward splitting**

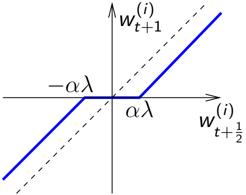
$$w_{t+\frac{1}{2}} = w_t - \alpha \nabla f(w_t)$$
$$w_{t+1} = \arg\min_{w} \left\{ \alpha \Psi(w) + \frac{1}{2} \left\| w - w_{t+\frac{1}{2}} \right\|_2^2 \right\}$$

or in compact form: $w_{t+1} = \mathbf{prox}_{\alpha\Psi}\big(w_t - \alpha \nabla f(w_t)\big)$

- $\Psi(w) = \lambda \|w\|_1$: **soft-thresholding**

$$w^{(i)}_{t+1} = \text{shrink}\Big(w^{(i)}_{t+\frac{1}{2}},\ \alpha\lambda\Big)$$

$$\text{shrink}(\omega, \alpha\lambda) = \begin{cases} \omega - \alpha\lambda & \text{if } \omega > \alpha\lambda \\ 0 & \text{if } |\omega| \le \alpha\lambda \\ \omega + \alpha\lambda & \text{if } \omega < -\alpha\lambda \end{cases}$$



Summary of background topics

- SGD for large scale learning
- online convex optimization (OCO)
- compressed sensing / sparse optimization
- proximal gradient method / soft thresholding

Summary of background topics

- SGD for large scale learning
- online convex optimization (OCO)
- compressed sensing / sparse optimization
- proximal gradient method / soft thresholding

the clash: put them together

SGD/OCO meets sparse optimization

regularized stochastic optimization:

$$\underset{w}{\text{minimize}} \quad \mathbf{E}_z f(w, z) + \Psi(w)$$

- $f(\cdot, z)$ convex for every z (e.g., least-squares, logistic regression)
- $\Psi(\cdot)$ convex and simple, especially $\Psi(w) = \lambda \|w\|_1$

SGD/OCO meets sparse optimization

regularized stochastic optimization:

$$\underset{w}{\text{minimize}} \quad \mathbf{E}_z f(w, z) + \Psi(w)$$

- $f(\cdot, z)$ convex for every z (e.g., least-squares, logistic regression)
- $\Psi(\cdot)$ convex and simple, especially $\Psi(w) = \lambda\|w\|_1$

stochastic subgradient method:

$$w_{t+1} = w_t - \alpha_t\big(g_t + \xi_t\big)$$

where $g_t = \nabla f(w_t, z_t)$, $\xi_t \in \partial\Psi(w_t)$, $\alpha_t \sim 1/\sqrt{t}$

SGD/OCO meets sparse optimization

regularized stochastic optimization:

$$\underset{w}{\text{minimize}} \quad \mathbf{E}_z f(w,z) + \Psi(w)$$

- $f(\cdot, z)$ convex for every z (e.g., least-squares, logistic regression)
- $\Psi(\cdot)$ convex and simple, especially $\Psi(w) = \lambda \|w\|_1$

stochastic subgradient method:

$$w_{t+1} = w_t - \alpha_t \big(g_t + \xi_t\big)$$

where $g_t = \nabla f(w_t, z_t)$, $\xi_t \in \partial \Psi(w_t)$, $\alpha_t \sim 1/\sqrt{t}$

sources of slow convergence rate $O(1/\sqrt{t})$:

- nonsmooth regularization $\Longrightarrow$ fix by proximal gradient method
- stochastic gradient $\Longrightarrow$ intrinsic, but sufficient for learning

SGD/OCO meets sparse optimization

regularized stochastic optimization:

$$\underset{w}{\text{minimize}} \quad \mathbf{E}_z f(w,z) + \Psi(w)$$

- $f(\cdot, z)$ convex for every z (e.g., least-squares, logistic regression)
- $\Psi(\cdot)$ convex and simple, especially $\Psi(w) = \lambda\|w\|_1$

stochastic subgradient method:

$$w_{t+1} = w_t - \alpha_t\big(g_t + \xi_t\big)$$

where $g_t = \nabla f(w_t, z_t)$, $\xi_t \in \partial\Psi(w_t)$, $\alpha_t \sim 1/\sqrt{t}$

sources of slow convergence rate $O(1/\sqrt{t})$:

- nonsmooth regularization $\Longrightarrow$ fix by proximal gradient method
- stochastic gradient $\Longrightarrow$ intrinsic, but sufficient for learning

what about sparsity?

Proximal SGD

$$w_{t+1} = \arg\min_{w} \left\{ f(w_t, z_t) + \langle g_t, w - w_t \rangle + \Psi(w) + \frac{1}{\alpha_t} \frac{\|w - w_t\|^2}{2} \right\}$$

- Duchi & Singer (2009 NIPS)
- Langford, Li & Zhang (2008 NIPS)
- Shalev-Shwartz & Tewari (2009)
- many others ...

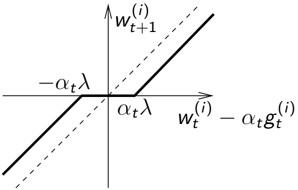
Proximal SGD

$$w_{t+1} = \arg\min_{w} \left\{ f(w_t, z_t) + \langle g_t, w - w_t \rangle + \Psi(w) + \frac{1}{\alpha_t} \frac{\|w - w_t\|^2}{2} \right\}$$

- Duchi & Singer (2009 NIPS)
- Langford, Li & Zhang (2008 NIPS)
- Shalev-Shwartz & Tewari (2009)
- many others ...

update for $\Psi(w) = \lambda\|w\|_1$:

$$\boxed{w_{t+1} = \mathrm{shrink}\big(w_t - \alpha_t g_t,\ \alpha_t \lambda\big)}$$



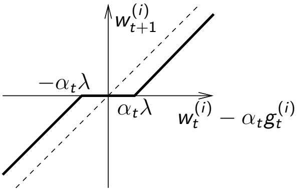
Proximal SGD

$$w_{t+1} = \arg\min_w \left\{ f(w_t, z_t) + \langle g_t, w - w_t \rangle + \Psi(w) + \frac{1}{\alpha_t} \frac{\|w - w_t\|^2}{2} \right\}$$

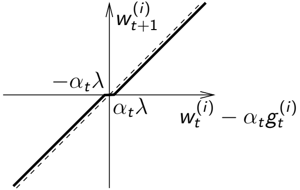
- Duchi & Singer (2009 NIPS)
- Langford, Li & Zhang (2008 NIPS)
- Shalev-Shwartz & Tewari (2009)
- many others ...

update for $\Psi(w) = \lambda \|w\|_1$:

$$\boxed{w_{t+1} = \text{shrink}\big(w_t - \alpha_t g_t,\ \alpha_t \lambda\big)}$$



$$\alpha_t \sim \frac{1}{\sqrt{t}} \to 0$$



Regularized dual averaging (RDA)

$$w_{t+1} = \arg\min_{w} \left\{ \frac{1}{t} \sum_{\tau=1}^{t} \Big[f(w_\tau, z_\tau) + \langle g_\tau, w - w_\tau \rangle \Big] + \Psi(w) + \frac{\gamma}{\sqrt{t}} \frac{\|w - w_0\|_2^2}{2} \right\}$$

Regularized dual averaging (RDA)

$$w_{t+1} = \arg\min_{w} \left\{ \frac{1}{t} \sum_{\tau=1}^{t} \Big[f(w_\tau, z_\tau) + \langle g_\tau, w - w_\tau \rangle \Big] + \Psi(w) + \frac{\gamma}{\sqrt{t}} \frac{\|w - w_0\|_2^2}{2} \right\}$$

$$= \arg\min_{w} \left\{ \langle \bar{g}_t, \, w \rangle + \Psi(w) + \frac{\gamma}{\sqrt{t}} \frac{\|w - w_0\|_2^2}{2} \right\}$$

where

$$\bar{g}_t = \frac{1}{t} \sum_{\tau=1}^{t} g_\tau$$

Regularized dual averaging (RDA)

$$w_{t+1} = \arg\min_w \left\{ \frac{1}{t}\sum_{\tau=1}^{t}\Big[f(w_\tau, z_\tau) + \langle g_\tau, w - w_\tau\rangle\Big] + \Psi(w) + \frac{\gamma}{\sqrt{t}}\frac{\|w - w_0\|_2^2}{2} \right\}$$

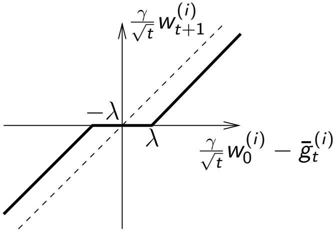
$$= \arg\min_w \left\{ \langle \bar{g}_t,\ w\rangle + \Psi(w) + \frac{\gamma}{\sqrt{t}}\frac{\|w - w_0\|_2^2}{2} \right\}$$

where

$$\bar{g}_t = \frac{1}{t}\sum_{\tau=1}^{t} g_\tau$$

- update for $\Psi(w) = \lambda\|w\|_1$

$$w_{t+1} = \frac{\sqrt{t}}{\gamma}\text{shrink}\left(\frac{\gamma}{\sqrt{t}}w_0 - \bar{g}_t,\ \lambda\right)$$



Regularized dual averaging (RDA)

$$w_{t+1} = \arg\min_w \left\{ \frac{1}{t}\sum_{\tau=1}^{t}\Big[f(w_\tau, z_\tau) + \langle g_\tau, w - w_\tau\rangle\Big] + \Psi(w) + \frac{\gamma}{\sqrt{t}}\frac{\|w - w_0\|_2^2}{2} \right\}$$

$$= \arg\min_w \left\{ \langle \bar{g}_t,\ w\rangle + \Psi(w) + \frac{\gamma}{\sqrt{t}}\frac{\|w - w_0\|_2^2}{2} \right\}$$

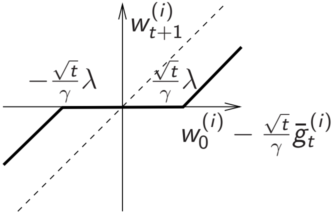
where

$$\bar{g}_t = \frac{1}{t}\sum_{\tau=1}^{t} g_\tau$$

- update for $\Psi(w) = \lambda\|w\|_1$

$$w_{t+1} = \frac{\sqrt{t}}{\gamma}\text{shrink}\left(\frac{\gamma}{\sqrt{t}}w_0 - \bar{g}_t,\ \lambda\right)$$

$$= \text{shrink}\left(w_0 - \frac{\sqrt{t}}{\gamma}\bar{g}_t,\ \frac{\sqrt{t}}{\gamma}\lambda\right)$$



Regularized dual averaging (RDA)

RDA: $$w_{t+1} = \arg\min_{w} \left\{ \langle \bar{g}_t, w \rangle + \Psi(w) + \frac{\beta_t}{t} \|w - w_0\|_2^2 \right\}$$

Regularized dual averaging (RDA)

RDA:

$$w_{t+1} = \arg\min_{w} \left\{ \langle \bar{g}_t, w \rangle + \Psi(w) + \frac{\beta_t}{t} \|w - w_0\|_2^2 \right\}$$

- **Nesterov's DA method:** $\Psi(w) = 0$ or indicator of convex set

Regularized dual averaging (RDA)

RDA: $$w_{t+1} = \arg\min_{w} \left\{ \langle \bar{g}_t, w \rangle + \Psi(w) + \frac{\beta_t}{t} \|w - w_0\|_2^2 \right\}$$

- **Nesterov's DA method:** $\Psi(w) = 0$ or indicator of convex set
- **online convex optimization:** minimize regret R_T

$$R_T := \sum_{t=1}^{T} \big(f_t(w_t) + \Psi(w_t) \big) - \min_{w \in \mathcal{S}} \sum_{t=1}^{T} \big(f_t(w) + \Psi(w) \big)$$

$R_T = O(\sqrt{T})$ or $O(\ln(T))$ if Ψ strongly convex

Regularized dual averaging (RDA)

RDA: $$w_{t+1} = \arg\min_{w} \left\{ \langle \bar{g}_t, w \rangle + \Psi(w) + \frac{\beta_t}{t} \|w - w_0\|_2^2 \right\}$$

- **Nesterov's DA method:** $\Psi(w) = 0$ or indicator of convex set
- **online convex optimization:** minimize regret R_T

$$R_T := \sum_{t=1}^{T} \big(f_t(w_t) + \Psi(w_t) \big) - \min_{w \in \mathcal{S}} \sum_{t=1}^{T} \big(f_t(w) + \Psi(w) \big)$$

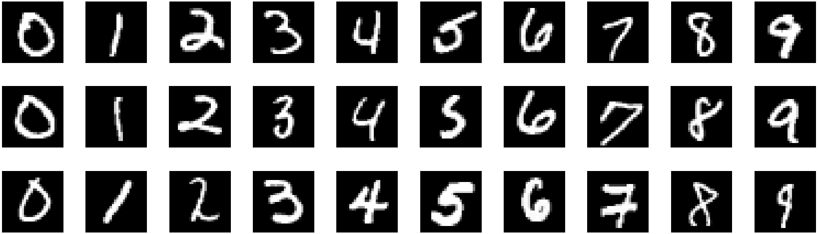
$R_T = O(\sqrt{T})$ or $O(\ln(T))$ if Ψ strongly convex

- **stochastic optimization:** (define $\bar{w}_t = \frac{1}{t} \sum_{\tau=1}^{t} w_\tau$)

$$\text{minimize} \quad \phi(w) := \mathbf{E}_z f(w, z) + \Psi(w)$$

$\phi(\bar{w}_t) - \phi_\star = O(1/\sqrt{t})$ or $O(\ln(t)/t)$ if Ψ strongly convex

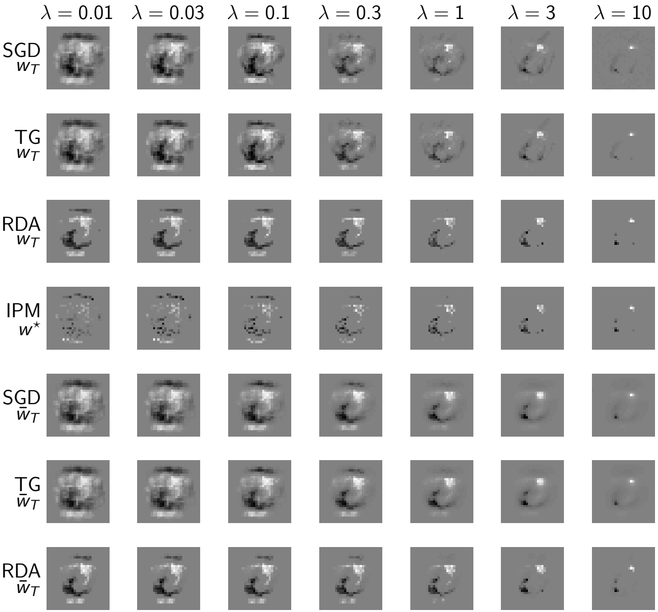
Experiments on MNIST



binary classification with logistic regression

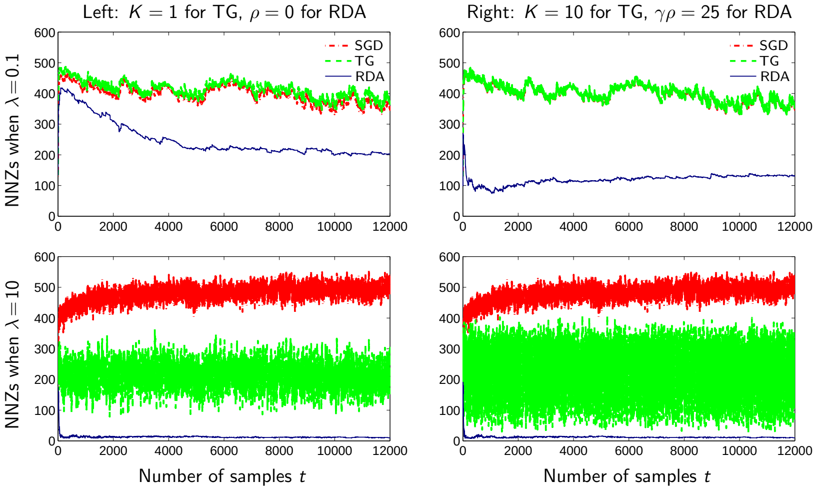
$$f(w,z) = \log\big(1+\exp\big(-y(\tilde{w}^T x + b)\big)\big), \qquad \Psi(w) = \lambda\|\tilde{w}\|_1$$

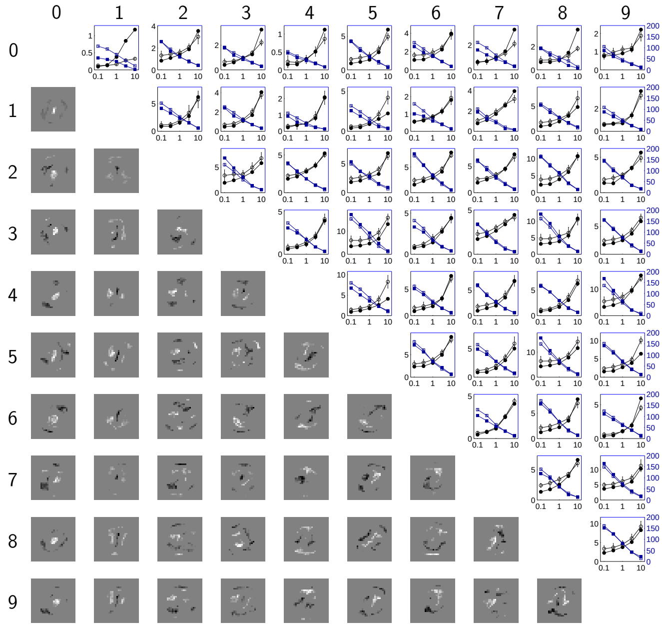
- $z = (x, y)$ where $x \in \mathbf{R}^{784}$ and $y \in \{+1, -1\}$
- $w = (\tilde{w}, b)$ where $\tilde{w} \in \mathbf{R}^{784}$ and $b \in \mathbf{R}$



$\lambda = 0.01$
$\lambda = 0.03$
$\lambda = 0.1$
$\lambda = 0.3$
$\lambda = 1$
$\lambda = 3$
$\lambda = 10$
SGD w_T
TG w_T
RDA w_T
IPM $w^\star$
SGD $\bar{w}_T$
TG $\bar{w}_T$
RDA $\bar{w}_T$

Sparsity in stochastic optimization




0
1
2
3
4
5
6
7
8
9

Interpretation and comparison

McMahan (2011)

- **RDA**

$$w_{t+1} = \arg\min_w \left\{ \frac{1}{t}\left\langle \sum_{\tau=1}^{t} g_\tau, w \right\rangle \qquad\qquad + \Psi(w) \qquad + \frac{\beta_t}{t}\|w - w_0\|_2^2 \right\}$$

Interpretation and comparison

McMahan (2011)

- **RDA**

$$w_{t+1} = \arg\min_w \left\{ \left\langle \sum_{\tau=1}^{t} g_\tau, w \right\rangle \qquad\qquad + t \cdot \Psi(w) \qquad + \beta_t \|w - w_0\|_2^2 \right\}$$

Interpretation and comparison

McMahan (2011)

- **RDA**

$$w_{t+1} = \arg\min_w \left\{ \left\langle \sum_{\tau=1}^{t} g_\tau, w \right\rangle \qquad\qquad + t \cdot \Psi(w) \qquad + \beta_t \|w - w_0\|_2^2 \right\}$$

- **Proximal SGD** (FOBOS)

$$w_{t+1} = \arg\min_w \left\{ \left\langle \sum_{\tau=1}^{t} g_\tau, w \right\rangle + \left\langle \sum_{\tau=1}^{t-1} \xi_\tau, w \right\rangle + \Psi(w) + \sum_{\tau=1}^{t} \eta_\tau \|w - w_\tau\|_2^2 \right\}$$

Interpretation and comparison

McMahan (2011)

- **RDA**

$$w_{t+1} = \arg\min_{w} \left\{ \left\langle \sum_{\tau=1}^{t} g_\tau, w \right\rangle \qquad + t \cdot \Psi(w) \qquad + \beta_t \|w - w_0\|_2^2 \right\}$$

- **Proximal SGD** (FOBOS)

$$w_{t+1} = \arg\min_{w} \left\{ \left\langle \sum_{\tau=1}^{t} g_\tau, w \right\rangle + \left\langle \sum_{\tau=1}^{t-1} \xi_\tau, w \right\rangle + \Psi(w) + \sum_{\tau=1}^{t} \eta_\tau \|w - w_\tau\|_2^2 \right\}$$

- **FTRL-Proximal** (McMahan 2011)

$$w_{t+1} = \arg\min_{w} \left\{ \left\langle \sum_{\tau=1}^{t} g_\tau, w \right\rangle \qquad + t \cdot \Psi(w) + \sum_{\tau=1}^{t} \eta_\tau \|w - w_\tau\|_2^2 \right\}$$

Further developments

- manifold identification (Lee and Wright 2012)
  - general convex case: $\mathbf{P}\big(w_t \in \mathcal{S}\big) \geq 1 - C_1 t^{-4}$
  - strongly convex case: $\mathbf{P}\big(w_t \in \mathcal{S}\big) \geq 1 - C_2 \sqrt{\ln(t)}\, t^{-2}$

  (under stronger assumptions and for sufficiently large t)

Further developments

- manifold identification (Lee and Wright 2012)
  - general convex case: $\mathbf{P}\big(w_t \in \mathcal{S}\big) \geq 1 - C_1 t^{-4}$
  - strongly convex case: $\mathbf{P}\big(w_t \in \mathcal{S}\big) \geq 1 - C_2 \sqrt{\ln(t)}\, t^{-2}$

  (under stronger assumptions and for sufficiently large t)

- accelerated RDA
  - X. (2010, longer version in JMLR)
  - Chen, Lin and Peña (2012)

Further developments

- manifold identification (Lee and Wright 2012)
  - general convex case: $\mathbf{P}\big(w_t \in \mathcal{S}\big) \geq 1 - C_1 t^{-4}$
  - strongly convex case: $\mathbf{P}\big(w_t \in \mathcal{S}\big) \geq 1 - C_2 \sqrt{\ln(t)}\, t^{-2}$

  (under stronger assumptions and for sufficiently large t)

- accelerated RDA
  - X. (2010, longer version in JMLR)
  - Chen, Lin and Peña (2012)

- RDA-ADMM (Suzuki 2013)

Further developments

- manifold identification (Lee and Wright 2012)
  - general convex case: $\mathbf{P}\big(w_t \in \mathcal{S}\big) \geq 1 - C_1 t^{-4}$
  - strongly convex case: $\mathbf{P}\big(w_t \in \mathcal{S}\big) \geq 1 - C_2 \sqrt{\ln(t)}\, t^{-2}$

  (under stronger assumptions and for sufficiently large t)

- accelerated RDA
  - X. (2010, longer version in JMLR)
  - Chen, Lin and Peña (2012)

- RDA-ADMM (Suzuki 2013)

- DA for distributed optimization (Duchi, Agarwal & Wainwright 2012)

- many others ...

Adaptive RDA

- AdaRDA (Duchi, Hazan & Singer 2011)

$$w_{t+1} = \arg\min_{w} \left\{ \frac{1}{t} \left\langle \sum_{\tau=1}^{t} g_\tau, w \right\rangle + \Psi(w) + \frac{1}{2t} w^T H_t w \right\}$$

where

$$H_t = \delta I + \mathrm{diag}(s_t), \qquad s_t^{(i)} = \sqrt{\sum_{\tau=1}^{t} \left(g_\tau^{(i)}\right)^2}$$

(proposed as one variant of the AdaGrad algorithm)

Adaptive RDA

- AdaRDA (Duchi, Hazan & Singer 2011)

$$w_{t+1} = \arg\min_{w} \left\{ \frac{1}{t} \left\langle \sum_{\tau=1}^{t} g_\tau, w \right\rangle + \Psi(w) + \frac{1}{2t} w^T H_t w \right\}$$

  where

$$H_t = \delta I + \mathrm{diag}(s_t), \qquad s_t^{(i)} = \sqrt{\sum_{\tau=1}^{t} \left(g_\tau^{(i)}\right)^2}$$

  (proposed as one variant of the AdaGrad algorithm)

- adaptive FTRL-Proximal (McMahan 2011)
- recent survey on adaptive online algorithms (MaMahan 2017)

Adaptive RDA

- AdaRDA (Duchi, Hazan & Singer 2011)

$$w_{t+1} = \arg\min_{w} \left\{ \frac{1}{t} \left\langle \sum_{\tau=1}^{t} g_\tau, w \right\rangle + \Psi(w) + \frac{1}{2t} w^T H_t w \right\}$$

where

$$H_t = \delta I + \mathrm{diag}(s_t), \qquad s_t^{(i)} = \sqrt{\sum_{\tau=1}^{t} \left(g_\tau^{(i)}\right)^2}$$

(proposed as one variant of the AdaGrad algorithm)

- adaptive FTRL-Proximal (McMahan 2011)
- recent survey on adaptive online algorithms (MaMahan 2017)

What about nonconvex optimization?

Nonconvex stochastic optimization

- **very active in recent years because of deep learning**
  - algorithms/theory well developed for generic problems
  - in theory, proximal SGD works with ℓ_1-regularization
  - may not obtain desired sparsity due to diminishing step sizes

Nonconvex stochastic optimization

- **very active in recent years because of deep learning**
  - algorithms/theory well developed for generic problems
  - in theory, proximal SGD works with ℓ_1-regularization
  - may not obtain desired sparsity due to diminishing step sizes
- **would variants of RDA work for nonconvex optimization?**

Nonconvex stochastic optimization

- **very active in recent years because of deep learning**
  - algorithms/theory well developed for generic problems
  - in theory, proximal SGD works with ℓ_1-regularization
  - may not obtain desired sparsity due to diminishing step sizes
- **would variants of RDA work for nonconvex optimization?**
  - promising for sparse CNN (Jia, Zhao, Zhang, He, Xu 2019)
  - need group Lasso to induce structured sparsity

Nonconvex stochastic optimization

- **very active in recent years because of deep learning**
  - algorithms/theory well developed for generic problems
  - in theory, proximal SGD works with ℓ_1-regularization
  - may not obtain desired sparsity due to diminishing step sizes
- **would variants of RDA work for nonconvex optimization?**
  - promising for sparse CNN (Jia, Zhao, Zhang, He, Xu 2019)
  - need group Lasso to induce structured sparsity
- **potential alternative: variance reduction techniques?**
  - proximal versions of SAG/SVRG/SAGA/SARAH/SPIDER
  - extensions to stochastic nonconvex optimization
  - provable convergence with constant step size: **better sparsity?**

Final reflections

motivation for RDA remain valid today:

- stochastic gradient and online algorithms are mainstays in ML
- (structured) sparsity is essential in scaling up large models

Final reflections

motivation for RDA remain valid today:

- stochastic gradient and online algorithms are mainstays in ML
- (structured) sparsity is essential in scaling up large models

additional challenges today:

- non-convexity
- more complex/structured models (deep neural networks)

Final reflections

motivation for RDA remain valid today:

- stochastic gradient and online algorithms are mainstays in ML
- (structured) sparsity is essential in scaling up large models

additional challenges today:

- non-convexity
- more complex/structured models (deep neural networks)

exciting progresses lie ahead