Stochastic Quasi-Gradient Methods: Variance Reduction via Jacobian Sketching

Robert M. Gower

Joint work with Peter Richtarik and Francis Bach

MODAL seminar 4th of September, 2018
Inria, Lille

Empirical Risk Minimization

$$\min_{w \in \mathbb{R}^d} f(w) \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^{n} f_i(w),$$

Datum functions	$f_i(w)$ is smooth and convex
Ridge Regression	$f_i(w) = (y^i - \langle w, x^i \rangle)^2 + \lambda \|\|w\|\|_2^2$
Conditional Random fields	$f_i(w) = -\ln\left(\frac{e^{w^\top F(x_i, y_i)}}{\sum_j e^{w^\top F(x_i, y_j)}}\right) + \lambda \|\|w\|\|_2^2$
Logistic regression	$f_i(w) = \ln(1 + e^{-y^i \langle w, x^i \rangle}) + \lambda \|\|w\|\|_2^2$

The Stochastic Gradient Method

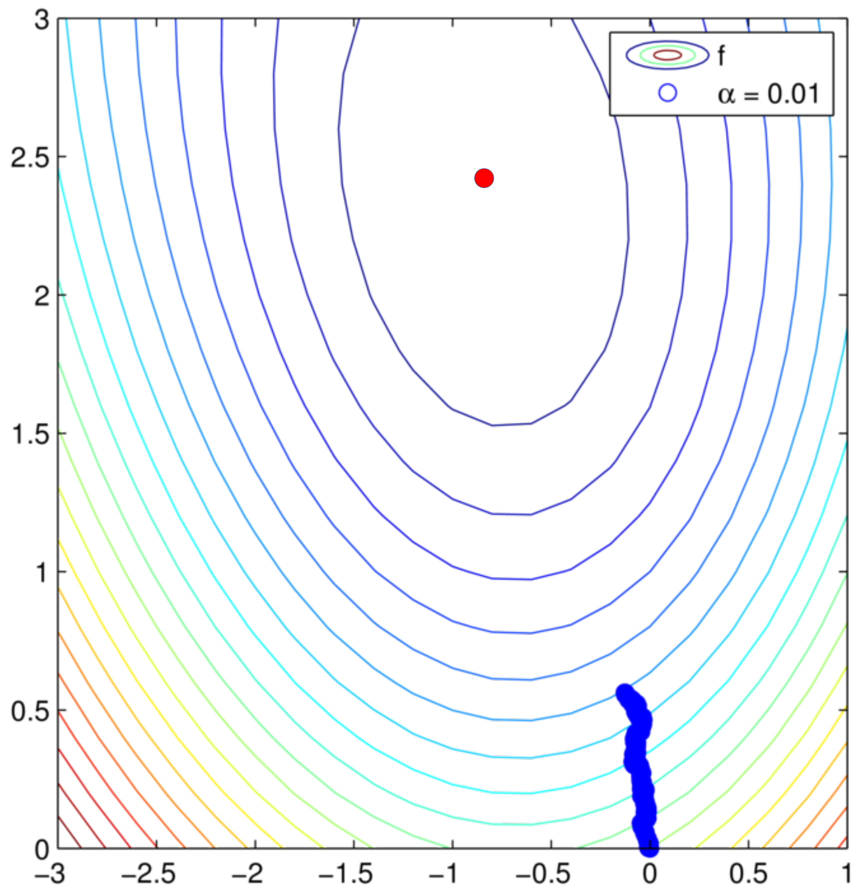
$$\min_{w\in\mathbb{R}^d} f(w) \overset{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n} f_i(w).$$

$$j \sim 1/n \quad \Rightarrow \quad \mathbb{E}_j[\nabla f_j(w)] \;=\; \frac{1}{n}\sum_{i=1}^{n}\nabla f_i(w) \;=\; \nabla f(w)$$

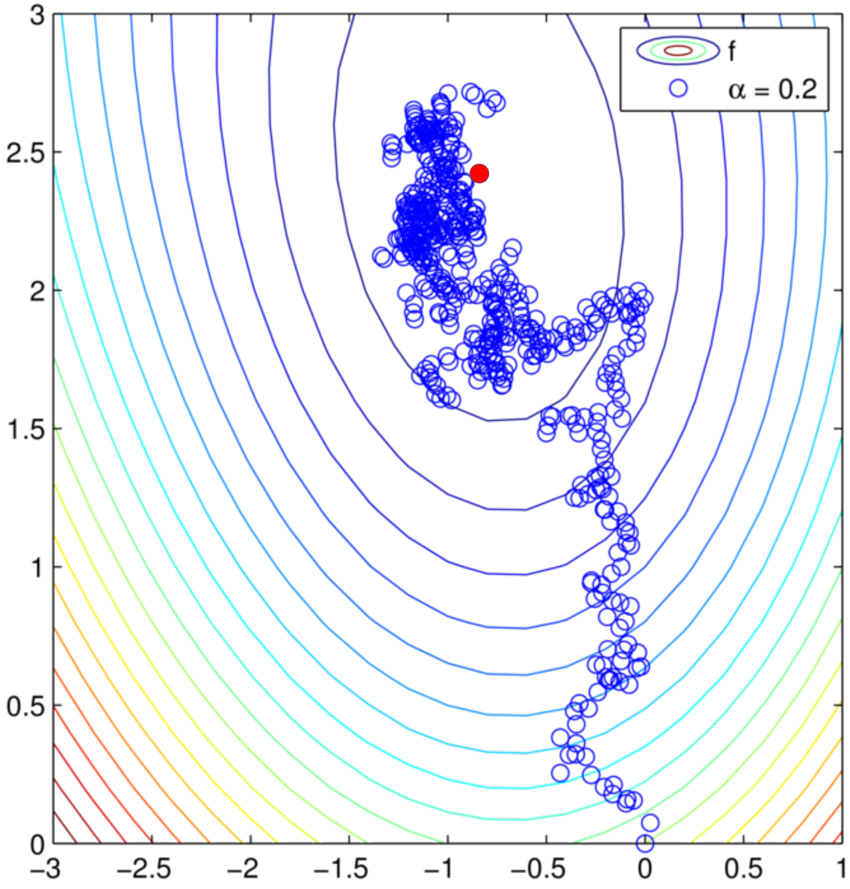
Stochastic Gradient Descent Algorithm

$$
\begin{array}{l}
\textbf{choose } \alpha_t = \alpha\, t^{-\beta},\ \alpha > 0, \beta > 0, \\
\textbf{for } t = 1, 2, 3, \ldots, T \\
\qquad \text{Sample } j \in \{1, \ldots, n\} \\
\qquad w^{t+1} = w^t - \alpha_t \nabla f_j(w^t) \\
\textbf{output } w^{T+1}
\end{array}
$$

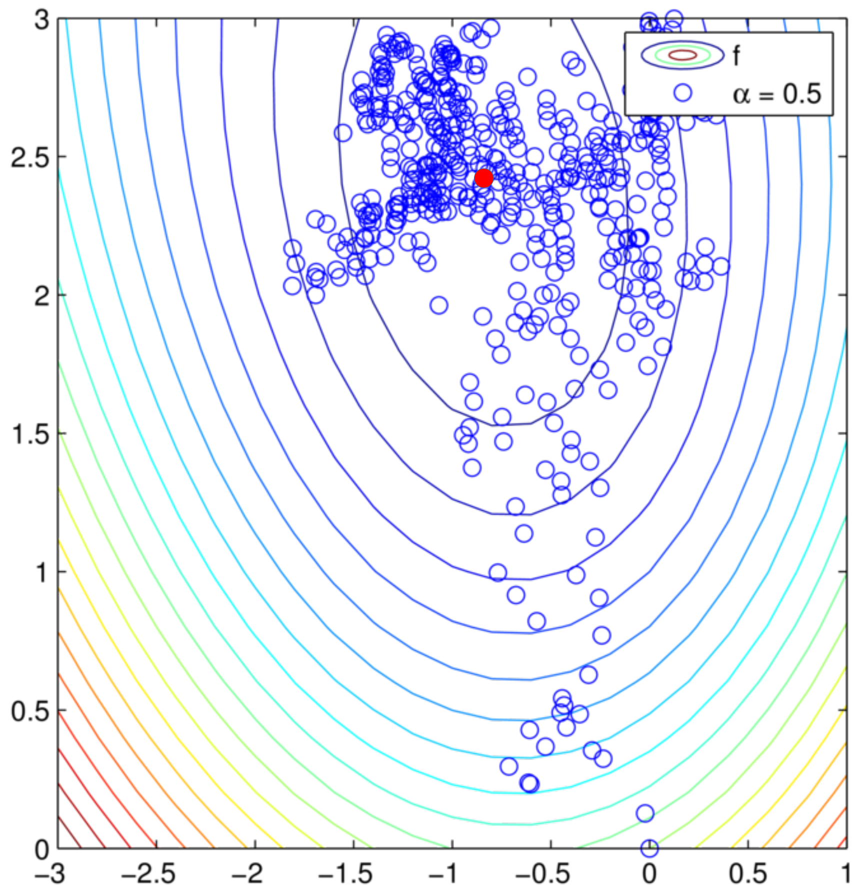
Stochastic Gradient Descent with small stepsizes



Stochastic Gradient Descent with large stepsizes



Stochastic Gradient Descent with larger stepsizes



Stoch. Grad Convergence

Theorem (Shrinking stepsize)

If $\alpha_t = \dfrac{1}{t\mu}$ then the iterates of the SGD method satisfy

$$\mathbb{E}\left[f(w^t) - f(w^*)\right] \leq O\left(\frac{1}{\mu t}\right)$$

Assuming

$$\mathbb{E}\left[||\nabla f_j(w^t)||_2^2\right] \leq B^2$$

Ohad Shamir and Tong Zhang (2013)
ICML, **Stochastic Gradient Descent for Non-smooth Optimization: Convergence Results and Optimal Averaging Schemes.**

Stoch. Grad Convergence

Theorem (Shrinking stepsize)

If $\alpha_t = \dfrac{1}{t\mu}$ then the iterates of the SGD method satisfy

$$\mathbb{E}\left[f(w^t) - f(w^*)\right] \leq O\left(\frac{1}{\mu t}\right)$$

Sublinear convergence

Assuming

$$\mathbb{E}\left[||\nabla f_j(w^t)||_2^2\right] \leq B^2$$

PDF Adobe

Ohad Shamir and Tong Zhang (2013)
ICML, **Stochastic Gradient Descent for Non-smooth Optimization: Convergence Results and Optimal Averaging Schemes.**

Building an estimate of the gradient

Mission Statement: Build an Estimate of the Gradient

Instead of using directly $\nabla f_j(w^t) \approx \nabla f(w^t)$
Use $\nabla f_j(w^t)$ to update estimate $g_t \approx \nabla f(w^t)$

Mission Statement: Build an Estimate of the Gradient

Instead of using directly $\nabla f_j(w^t) \approx \nabla f(w^t)$
Use $\nabla f_j(w^t)$ to update estimate $g_t \approx \nabla f(w^t)$

$$w^{t+1} = w^t - \alpha g^t$$

Mission Statement: Build an Estimate of the Gradient

Instead of using directly $\nabla f_j(w^t) \approx \nabla f(w^t)$
Use $\nabla f_j(w^t)$ to update estimate $g_t \approx \nabla f(w^t)$

$$w^{t+1} = w^t - \alpha g^t$$

We would like gradient estimate such that:

Unbiased

$$\mathbb{E}[g^t] = \nabla f(w^t)$$

Converges in *L2*

$$\mathbb{E}[||g^t - \nabla f(w^t)||_2^2] \underset{w^t \to w^*}{\longrightarrow} 0$$

Mission Statement: Build an Estimate of the Gradient

Instead of using directly $\nabla f_j(w^t) \approx \nabla f(w^t)$
Use $\nabla f_j(w^t)$ to update estimate $g_t \approx \nabla f(w^t)$

$$w^{t+1} = w^t - \alpha g^t$$

We would like gradient estimate such that:

No need to assume
$\mathbb{E}\left[||\nabla f_j(w^t)||_2^2\right] \leq B^2$

Unbiased

$$\mathbb{E}[g^t] = \nabla f(w^t)$$

Converges in $L2$

$$\mathbb{E}[||g^t - \nabla f(w^t)||_2^2] \underset{w^t \rightarrow w^*}{\longrightarrow} 0$$

Build an Estimate of a Jacobian

$$\min_{w \in \mathbf{R}^d} f(w) \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^{n} f_i(w)$$

$$F(w) \stackrel{\text{def}}{=} (f_1(w), \ldots, f_n(w))$$

$$DF(w) = (\nabla f_1(w), \ldots, \nabla f_n(w)) \in \mathbb{R}^{d \times n}$$

Build an Estimate of a Jacobian

$$\min_{w \in \mathbf{R}^d} f(w) \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n} f_i(w)$$

$$F(w) \stackrel{\text{def}}{=} (f_1(w), \ldots, f_n(w))$$

$$DF(w) = (\nabla f_1(w), \ldots, \nabla f_n(w)) \in \mathbb{R}^{d \times n}$$

$$\nabla f(w) = \frac{1}{n} DF(w)\mathbf{1}, \quad \text{where } \mathbf{1}^\top = (1, 1, \ldots, 1) \in \mathbf{R}^n$$

Build an Estimate of a Jacobian

$$\min_{w \in \mathbf{R}^d} f(w) \overset{\text{def}}{=} \frac{1}{n} \sum_{i=1}^{n} f_i(w)$$

$$F(w) \overset{\text{def}}{=} (f_1(w), \ldots, f_n(w))$$

$$DF(w) = (\nabla f_1(w), \ldots, \nabla f_n(w)) \in \mathbb{R}^{d \times n}$$

$$\nabla f(w) = \frac{1}{n} DF(w)\mathbf{1}, \quad \text{where } \mathbf{1}^\top = (1, 1, \ldots, 1) \in \mathbf{R}^n$$

Build an Estimate of a Jacobian

$$\min_{w\in\mathbf{R}^d} f(w) \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=1}^n f_i(w)$$

$$F(w) \stackrel{\text{def}}{=} (f_1(w), \ldots, f_n(w))$$

$$DF(w) = (\nabla f_1(w), \ldots, \nabla f_n(w)) \in \mathbb{R}^{d\times n}$$

$$\nabla f(w) = \frac{1}{n} DF(w)\mathbf{1}, \quad \text{where } \mathbf{1}^\top = (1, 1, \ldots, 1) \in \mathbf{R}^n$$

$$g^t = \frac{1}{n} J^t \mathbf{1}, \quad \text{where} \quad J^t \approx DF(w^t) \in \mathbb{R}^{d\times n}$$

Updating the Jacobian Estimate

Assume: Only have access to $DF(w^t)e_j = \nabla f_j(w^t)$

$$J \quad = \quad DF(w^t)$$

Updating the Jacobian Estimate

Assume: Only have access to $DF(w^t)e_j = \nabla f_j(w^t)$

$$J\, e_j = \quad DF(w^t)e_j, \quad j \sim \mathcal{U}\{1, n\}$$

Updating the Jacobian Estimate

Assume: Only have access to $DF(w^t)e_j = \nabla f_j(w^t)$

Sketch and Project the Jacobian

$$J^t = \arg\min_{J \in \mathbb{R}^{d\times n}} ||J - J^{t-1}||^2$$

$$J\, e_j = \quad DF(w^t)e_j, \quad j \sim \mathcal{U}\{1,n\}$$

Updating the Jacobian Estimate

Assume: Only have access to $DF(w^t)e_j = \nabla f_j(w^t)$

Sketch and Project the Jacobian

$$J^t = \arg\min_{J \in \mathbb{R}^{d \times n}} ||J - J^{t-1}||^2$$

$$J e_j = \quad DF(w^t)e_j, \quad j \sim \mathcal{U}\{1, n\}$$

Solution: $J^t = J^{t-1} + (J^{t-1} - DF(w^t))e_j e_j^\top$

Updating the Jacobian Estimate

Assume: Only have access to $DF(w^t)e_j = \nabla f_j(w^t)$

Any other ways to cheaply extract information from $DF(w^t)$?

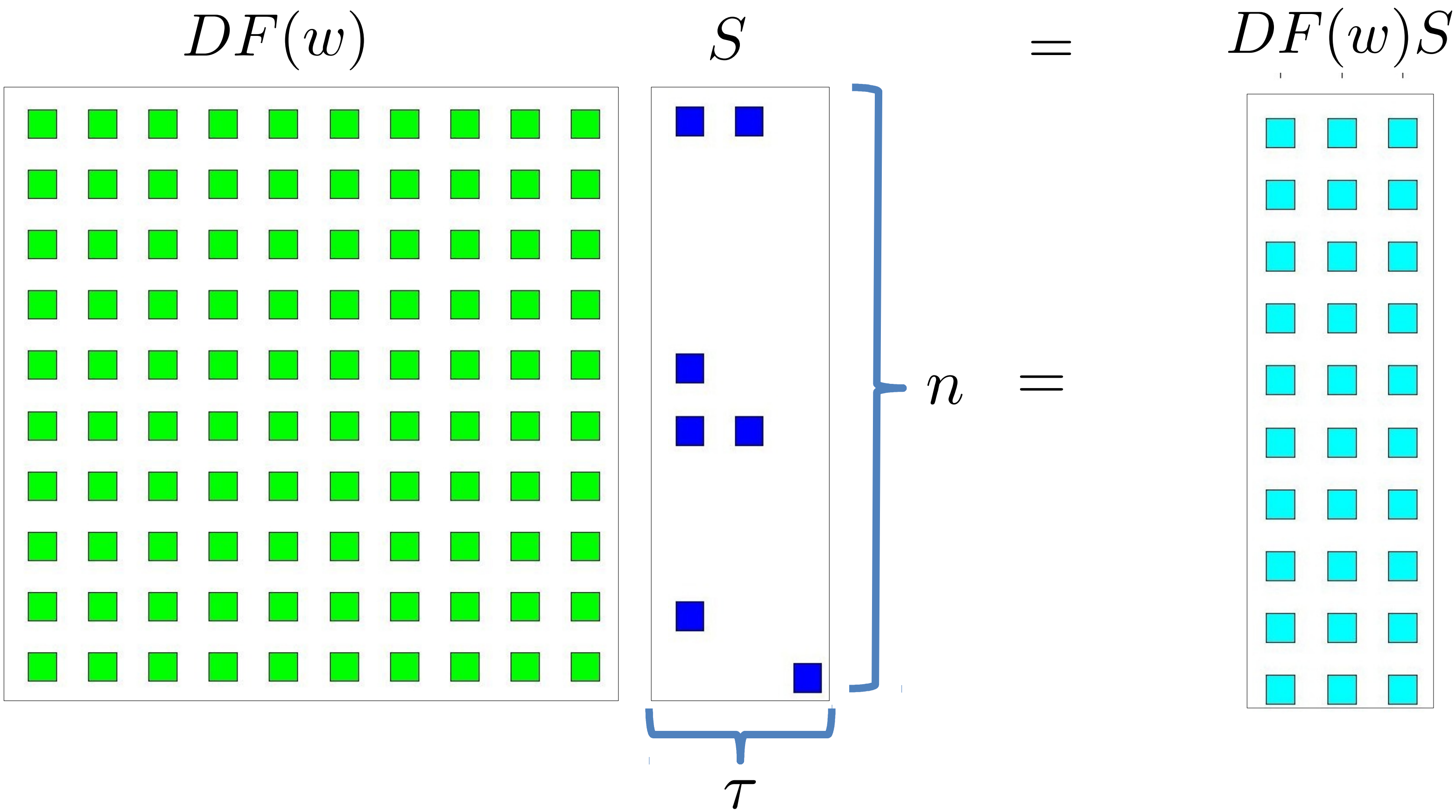
Sketch and Project the Jacobian

$$J^t = \arg\min_{J \in \mathbb{R}^{d\times n}} ||J - J^{t-1}||^2$$

$$J\,e_j = \quad DF(w^t)e_j, \quad j \sim \mathcal{U}\{1, n\}$$

Solution: $J^t = J^{t-1} + (J^{t-1} - DF(w^t))e_j e_j^\top$

Stochastic Sparse Sketches



Sparse Stochastic Matrix

$S \sim \mathcal{D}$ fixed distribution $S \in \mathbb{R}^{n \times \tau}$ a sparse matrix and $\tau \ll d, n$

David P. Woodruff (2014), Foundations and Trends® in Theoretical Computer, **Sketching as a Tool for Numerical Linear Algebra.**

Stochastic Sparse Sketches

SGD Sketch:

$$S = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \end{pmatrix} = e_j$$

$$DF(w)S = \nabla f_j(w)$$

Averaging Sketch:

$$S = \begin{pmatrix} a_1 \\ 0 \\ a_3 \\ a_4 \end{pmatrix} = \sum_{i \in C} a_i e_i$$

$$DF(w)S = \sum_{i \in C} a_i \nabla f_i(w)$$

Mini-batch Sketch:

$$S = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} = I_C$$

$$DF(w)S = \sum_{i \in C} \nabla f_i(w) e_i^\top$$

Sketch and project the Jacobian

$$J^{t+1} \quad = \quad \arg \min_{J \in \mathbb{R}^{d \times n}} ||J - J^t||^2$$
$$\text{subject to } JS = DF(w^t)S$$

$$J^{t+1} = J^t - (J^t - DF(w^t))S(S^\top S)^{-1}S^\top$$

RMG, P. Richtarik (2015), SIAM. J. Matrix Anal. & Appl.
Randomized Iterative Methods for Linear Systems

Sketch and project the Jacobian

$$J^{t+1} = \arg\min_{J \in \mathbb{R}^{d\times n}} ||J - J^t||^2 \quad \text{subject to } JS = DF(w^t)S$$

$$J^{t+1} = J^t - (J^t - DF(w^t))S(S^\top S)^{-1}S^\top$$

$$g^t = \frac{\theta}{n} J^{t+1}\mathbf{1} + \frac{1-\theta}{n} J^t \mathbf{1}$$

RMG, P. Richtarik (2015), SIAM. J. Matrix Anal. & Appl.
Randomized Iterative Methods for Linear Systems

Sketch and project the Jacobian

$$
\begin{aligned}
J^{t+1} \quad &= \quad \arg\min_{J \in \mathbb{R}^{d\times n}} \|J - J^t\|^2 \\
&\quad\quad \text{subject to } JS = DF(w^t)S
\end{aligned}
$$

$$J^{t+1} = J^t - (J^t - DF(w^t))S(S^\top S)^{-1}S^\top$$

$\theta = 1 \Rightarrow$ low variance but biased

$$g^t = \frac{\theta}{n} J^{t+1}\mathbf{1} + \frac{1-\theta}{n} J^t \mathbf{1}$$

RMG, P. Richtarik (2015), SIAM. J. Matrix Anal. & Appl.
Randomized Iterative Methods for Linear Systems

Sketch and project the Jacobian

$$\|J\|_W^2 = \mathbf{Tr}(J^\top J W)$$

$$\begin{aligned} J^{t+1} \quad = \quad & \arg\min_{J \in \mathbb{R}^{d\times n}} \|J - J^t\|_{W}^2 \\ & \text{subject to } JS = DF(w^t)S \end{aligned}$$

$$J^{t+1} = J^t - (J^t - DF(w^t))S(S^\top W^{-1} S)^{-1} S^\top W^{-1}$$

$$g^t = \frac{\theta}{n} J^{t+1}\mathbf{1} + \frac{1-\theta}{n} J^t \mathbf{1}$$

PDF Adobe

RMG, P. Richtarik (2015), SIAM. J. Matrix Anal. & Appl.
Randomized Iterative Methods for Linear Systems

Sketch and project the Jacobian

$$J^{t+1} \quad = \quad \arg\min_{J \in \mathbb{R}^{d\times n}} ||J - J^t||^2$$
$$\text{subject to } JS = DF(w^t)S$$

$$J^{t+1} = J^t - (J^t - DF(w^t))S(S^\top S)^{-1}S^\top$$

$$g^t = \frac{\theta}{n} J^{t+1}\mathbf{1} + \frac{1-\theta}{n} J^t \mathbf{1}$$

RMG, P. Richtarik (2015), SIAM. J. Matrix Anal. & Appl.
Randomized Iterative Methods for Linear Systems

Sketch and project the Jacobian

$$J^{t+1} = \arg\min_{J \in \mathbb{R}^{d \times n}} ||J - J^t||^2$$
$$\text{subject to } JS = DF(w^t)S$$

$$J^{t+1} = J^t - (J^t - DF(w^t))\underbrace{S(S^\top S)^{-1}S^\top}_{=: P_S}$$

$$g^t = \frac{\theta}{n} J^{t+1}\mathbf{1} + \frac{1-\theta}{n} J^t \mathbf{1}$$

PDF Adobe

RMG, P. Richtarik (2015), SIAM. J. Matrix Anal. & Appl.

Randomized Iterative Methods for Linear Systems

Unbiased Condition

Lemma. If $\mathbb{E}_S[P_S]\mathbf{1} = \frac{1}{\theta}\mathbf{1}$ then

$$\mathbb{E}_S[g^t] = \nabla f(w^t)$$

consequently g^t is an unbiased estimator.

Proof:

$$\begin{aligned}
\mathbb{E}_S[g^t] &= \frac{1}{n}J^{t-1}\mathbf{1} - \frac{\theta}{n}(J^{t-1} - DF(w^t))\mathbb{E}_S[S(S^\top S)^{-1}S^\top]\mathbf{1} \\
&= \frac{1}{n}J^{t-1}\mathbf{1} - \frac{\theta}{n\theta}(J^{t-1} - DF(w^t))\mathbf{1} \\
&= \frac{1}{n}J^{t-1}\mathbf{1} - \frac{1}{n}J^{t-1}\mathbf{1} + \frac{1}{n}DF(w^t))\mathbf{1} \quad = \quad \nabla f(w^t)
\end{aligned}$$

Unbiased Condition

Lemma. If $\mathbb{E}_S[P_S]\mathbf{1} = \frac{1}{\theta}\mathbf{1}$ then

$$\mathbb{E}_S[g^t] = \nabla f(w^t)$$

consequently g^t is an unbiased estimator.

Proof:

$$\begin{aligned}
\mathbb{E}_S[g^t] &= \frac{1}{n}J^{t-1}\mathbf{1} - \frac{\theta}{n}(J^{t-1} - DF(w^t))\mathbb{E}_S[S(S^\top S)^{-1}S^\top]\mathbf{1} \\
&= \frac{1}{n}J^{t-1}\mathbf{1} - \frac{\theta}{n\theta}(J^{t-1} - DF(w^t))\mathbf{1} \\
&= \cancel{\frac{1}{n}J^{t-1}}\mathbf{1} - \cancel{\frac{1}{n}J^{t-1}}\mathbf{1} + \frac{1}{n}DF(w^t))\mathbf{1} \quad = \quad \nabla f(w^t)
\end{aligned}$$

Exercise

Let $\mathbb{P}[S = e_i] = \dfrac{1}{n}$ for $i = 1, \ldots, n$. Show that

$$\mathbb{E}[P_S]\mathbf{1} = \mathbb{E}[S(S^\top S)^{-1} S^\top]\mathbf{1} = \frac{1}{n}\mathbf{1}$$

Proof:

Exercise

Let $\mathbb{P}[S = e_i] = \frac{1}{n}$ for $i = 1, \ldots, n$. Show that

$$\mathbb{E}[P_S]\mathbf{1} = \mathbb{E}[S(S^\top S)^{-1}S^\top]\mathbf{1} = \frac{1}{n}\mathbf{1}$$

Proof:

Exercise

Let $\mathbb{P}[S = e_i] = \dfrac{1}{n}$ for $i = 1, \ldots, n$. Show that

$$\boxed{\mathbb{E}[P_S]\mathbf{1} = \mathbb{E}[S(S^\top S)^{-1}S^\top]\mathbf{1} = \frac{1}{n}\mathbf{1}}$$

Proof:

$$\begin{aligned}
\mathbb{E}[S(S^\top S)^{-1}S^\top]\mathbf{1} &= \textstyle\sum_{i=1}^n \frac{1}{n}\frac{e_i e_i^\top}{e_i^\top e_i} \\
&= \frac{1}{n}\sum_{i=1}^n e_i e_i^\top \mathbf{1} \\
&= \frac{1}{n} I \mathbf{1} = \frac{1}{n}\mathbf{1}
\end{aligned}$$

Exercise

$$\mathbb{E}_S[P_S]\mathbf{1} = \frac{1}{\theta}\mathbf{1}$$

Let $\mathbb{P}[S = e_i] = \dfrac{1}{n}$ for $i = 1, \ldots, n$. Show that

$$\mathbb{E}[P_S]\mathbf{1} = \mathbb{E}[S(S^\top S)^{-1}S^\top]\mathbf{1} = \frac{1}{n}\mathbf{1}$$

Proof:

$$\begin{aligned}
\mathbb{E}[S(S^\top S)^{-1}S^\top]\mathbf{1} &= \textstyle\sum_{i=1}^{n} \frac{1}{n} \frac{e_i e_i^\top}{e_i^\top e_i} \\
&= \frac{1}{n}\sum_{i=1}^{n} e_i e_i^\top \mathbf{1} \\
&= \frac{1}{n} I \mathbf{1} = \frac{1}{n}\mathbf{1}
\end{aligned}$$

Exercise

$$\mathbb{E}_S[P_S]\mathbf{1} = \frac{1}{\theta}\mathbf{1}$$

Let $\mathbb{P}[S = e_i] = \dfrac{1}{n}$ for $i = 1, \ldots, n$. Show that

$$\mathbb{E}[P_S]\mathbf{1} = \mathbb{E}[S(S^\top S)^{-1}S^\top]\mathbf{1} = \frac{1}{n}\mathbf{1}$$

$$\theta = n$$

Proof:

$$\begin{aligned}
\mathbb{E}[S(S^\top S)^{-1}S^\top]\mathbf{1} &= \textstyle\sum_{i=1}^n \frac{1}{n}\frac{e_i e_i^\top}{e_i^\top e_i} \\
&= \frac{1}{n}\sum_{i=1}^n e_i e_i^\top \mathbf{1} \\
&= \frac{1}{n} I \mathbf{1} = \frac{1}{n}\mathbf{1}
\end{aligned}$$

A Jacobian Based Method

JacSketch Algorithm

choose distribution $\mathcal{D}$ and the bias-correcting variable $\theta > 0$
choose $\alpha > 0, w^1 \in \mathbb{R}^d, J^1 \in \mathbb{R}^{d\times n}$
for $t = 1, \ldots, T$
  sample $S \sim \mathcal{D}$
  calculate sketch $DF(w^t)S$
  update $J^{t+1} = J^t - (J^t - DF(w^t))S(S^\top S)^{-1}S^\top$
  calculate $g^t = \frac{\theta}{n}J^{t+1}\mathbf{1} + \frac{1-\theta}{n}J^t\mathbf{1}$
  step $w^{t+1} = w^t - \alpha g^t$

EXE: Show that if S is invertible, this algorithm is gradient descent

A Jacobian Based Method

JacSketch Algorithm

choose distribution $\mathcal{D}$ and the bias-correcting variable $\theta > 0$
choose $\alpha > 0, w^1 \in \mathbb{R}^d, J^1 \in \mathbb{R}^{d \times n}$
for $t = 1, \ldots, T$
  sample $S \sim \mathcal{D}$
  calculate sketch $DF(w^t)S$
  update $J^{t+1} = J^t - (J^t - DF(w^t))S(S^\top S)^{-1}S^\top$
  calculate $g^t = \frac{\theta}{n} J^{t+1}\mathbf{1} + \frac{1-\theta}{n} J^t \mathbf{1}$
  step $w^{t+1} = w^t - \alpha g^t$

Looks expensive and complicated. Investigate

EXE: Show that if S is invertible, this algorithm is gradient descent

Example: minibatch-SAGA

$$\mathbb{P}[S = I_C] = 1/\binom{n}{\tau}, \text{ for all } C \subset \{1, \ldots, n\} \text{ with } |C| = \tau$$

Example: minibatch-SAGA

$$\mathbb{P}[S = I_C] = 1/\binom{n}{\tau}, \text{ for all } C \subset \{1,\ldots,n\} \text{ with } |C| = \tau$$

Jacobian update

$$J_j^{t+1} = \begin{cases} \nabla f_j(w^t) & \text{if } j \in C, \\ J_j^t & \text{if } j \neq C. \end{cases}$$

Gradient estimate

$$g^t = \frac{1}{n} J^t \mathbf{1} - \frac{1}{\tau} \sum_{j \in C} (J_j^t - \nabla f_j(w^t))$$

Example: minibatch-SAGA

$$\mathbb{P}[S = I_C] = 1/\binom{n}{\tau}, \text{ for all } C \subset \{1,\ldots,n\} \text{ with } |C| = \tau$$

Jacobian update

$$J_j^{t+1} = \begin{cases} \nabla f_j(w^t) & \text{if } j \in C, \\ J_j^t & \text{if } j \neq C. \end{cases}$$

Gradient estimate

$$g^t = \frac{1}{n} J^t \mathbf{1} - \frac{1}{\tau} \sum_{j \in C} (J_j^t - \nabla f_j(w^t))$$

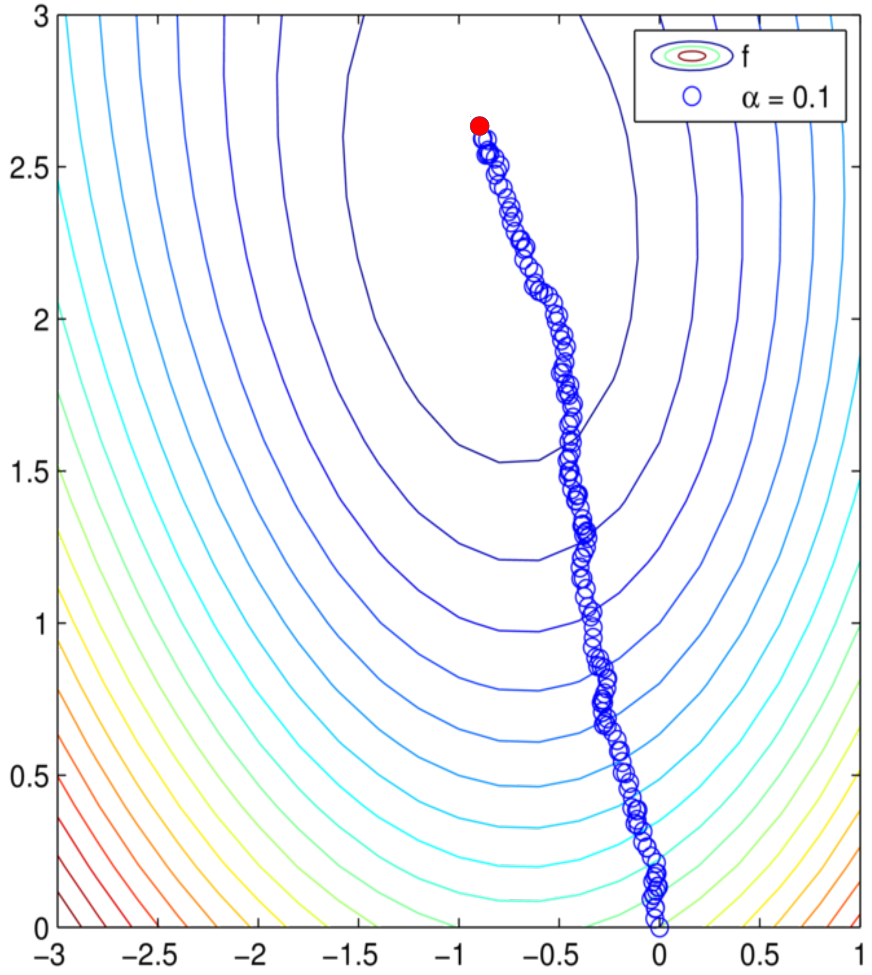
Unbiased Condition:

$$\mathbb{E}_S[P_S]\mathbf{1} = \frac{\tau}{n}\mathbf{1}$$

$$\theta = \frac{n}{\tau}$$

M. Schmidt, N. Le Roux, F. Bach (2016), Mathematical Programming
Minimizing Finite Sums with the Stochastic Average Gradient.

Example: The Stochastic Average Gradient (SAGA)



Over Halfway point query

Option I: See reformulation of method as a weird version of SGD (new interpretation)

Option II: See proof of convergence (new proofs)

Over Halfway point query

Option I: See reformulation of method as a weird version of SGD (new interpretation)

Option II: See proof of convergence (new proofs)

Stochastic Gradient Descent Applied to Stochastic Reformulation Viewpoint

Simple Stochastic Reformulation

Unbiased condition

$$\min_{w \in \mathbf{R}^d} f(w) := \mathbb{E}_{S\sim\mathcal{D}}\left[f_S(w)\right] \qquad (\mathbf{SP})$$

where $\quad f_S(w) := \dfrac{\theta}{n}\langle F(w), P_S \mathbf{1}\rangle, \qquad$ and $\quad \mathbb{E}_{S\sim\mathcal{D}}[P_S]\mathbf{1} = \frac{1}{\theta}\mathbf{1}$

Simple Stochastic Reformulation

Unbiased condition

$$\min_{w \in \mathbf{R}^d} f(w) := \mathbb{E}_{S \sim \mathcal{D}}\left[f_S(w)\right] \qquad (\mathbf{SP})$$

where $\quad f_S(w) := \dfrac{\theta}{n}\langle F(w), P_S \mathbf{1}\rangle, \qquad$ and $\quad \mathbb{E}_{S\sim\mathcal{D}}[P_S]\mathbf{1} = \frac{1}{\theta}\mathbf{1}$

Lemma. Solving **SP** is equivalent to solving our original problem

Simple Stochastic Reformulation

Unbiased condition

$$\min_{w\in\mathbf{R}^d} f(w) := \mathbb{E}_{S\sim\mathcal{D}}\left[f_S(w)\right] \qquad (\mathbf{SP})$$

where $\quad f_S(w) := \dfrac{\theta}{n}\langle F(w), P_S\mathbf{1}\rangle, \qquad$ and $\quad \mathbb{E}_{S\sim\mathcal{D}}[P_S]\mathbf{1} = \frac{1}{\theta}\mathbf{1}$

Lemma. Solving **SP** is equivalent to solving our original problem

Proof: $\mathbb{E}_{S\sim\mathcal{D}}\left[\dfrac{\theta}{n}\langle F(w), P_S\mathbf{1}\rangle\right] = \dfrac{\theta}{n}\langle F(x), \mathbb{E}[P_S]\mathbf{1}\rangle = \dfrac{1}{n}\langle F(w), \mathbf{1}\rangle = \dfrac{1}{n}\displaystyle\sum_{i=1}^n f_i(w)$

Unbiased condition

Simple Stochastic Reformulation

$$\min_{w \in \mathbf{R}^d} f(w) := \mathbb{E}_{S\sim\mathcal{D}}\left[f_S(w)\right] \qquad (\mathbf{SP})$$

Unbiased condition

where $\quad f_S(w) := \dfrac{\theta}{n}\langle F(w), P_S \mathbf{1}\rangle, \qquad$ and $\quad \mathbb{E}_{S\sim\mathcal{D}}[P_S]\mathbf{1} = \frac{1}{\theta}\mathbf{1}$

Lemma. Solving **SP** is equivalent to solving our original problem

Proof: $\mathbb{E}_{S\sim\mathcal{D}}\left[\dfrac{\theta}{n}\langle F(w), P_S\mathbf{1}\rangle\right] = \dfrac{\theta}{n}\langle F(x), \mathbb{E}[P_S]\mathbf{1}\rangle = \dfrac{1}{n}\langle F(w), \mathbf{1}\rangle = \dfrac{1}{n}\sum_{i=1}^{n} f_i(w)$

Unbiased condition

Apply SGD to solve **SP**?

Simple Stochastic Reformulation

Unbiased condition

$$\min_{w\in\mathbf{R}^d} f(w) := \mathbb{E}_{S\sim\mathcal{D}}\left[f_S(w)\right] \qquad (\mathbf{SP})$$

where $\quad f_S(w) := \dfrac{\theta}{n}\langle F(w), P_S\mathbf{1}\rangle, \qquad$ and $\quad \mathbb{E}_{S\sim\mathcal{D}}[P_S]\mathbf{1} = \frac{1}{\theta}\mathbf{1}$

Lemma. Solving **SP** is equivalent to solving our original problem

Proof: $\mathbb{E}_{S\sim\mathcal{D}}\left[\dfrac{\theta}{n}\langle F(w), P_S\mathbf{1}\rangle\right] = \dfrac{\theta}{n}\langle F(x), \mathbb{E}[P_S]\mathbf{1}\rangle = \dfrac{1}{n}\langle F(w), \mathbf{1}\rangle = \dfrac{1}{n}\sum_{i=1}^n f_i(w)$

Unbiased condition

Apply SGD to solve **SP**?

Sample $S \sim \mathcal{D}$

$$w^{t+1} = w^t - \alpha\nabla f_S(w^t) = w^t - \alpha\frac{\theta}{n}DF(w)P_S\mathbf{1}$$

Controlled Stochastic Reformulation

Let $J \in \mathbb{R}^{d\times n}$.

Control Variate: $\psi_{S,J}(w) := \frac{1}{n}\langle J^\top w, (I - \theta P_S)\mathbf{1}\rangle$

$$\text{Unbiased Condition} \quad \Rightarrow \quad (I - \theta \mathbb{E}_{S\sim\mathcal{D}}[P_S])\mathbf{1} = 0$$

$$\Rightarrow \quad \mathbb{E}_{S\sim\mathcal{D}}[\psi_{S,J}(w)] = 0$$

Controlled Stochastic Reformulation

Let $J \in \mathbb{R}^{d\times n}$.

Control Variate: $\psi_{S,J}(w) := \frac{1}{n}\langle J^\top w, (I - \theta P_S)\mathbf{1}\rangle$

$$\text{Unbiased Condition} \quad \Rightarrow \quad (I - \theta \mathbb{E}_{S\sim\mathcal{D}}[P_S])\mathbf{1} = 0$$

$$\Rightarrow \quad \mathbb{E}_{S\sim\mathcal{D}}[\psi_{S,J}(w)] = 0$$

$$\min_{w\in\mathbf{R}^d} \mathbb{E}_{S\sim\mathcal{D}}\left[f_{S,J}(w) := f_S(w) + \psi_{S,J}(w)\right], \qquad (\mathbf{CSP})$$

Controlled Stochastic Reformulation

Let $J \in \mathbb{R}^{d\times n}$.

Control Variate: $\psi_{S,J}(w) := \frac{1}{n}\langle J^\top w, (I-\theta P_S)\mathbf{1}\rangle$

$$\text{Unbiased Condition} \quad \Rightarrow \quad (I - \theta \mathbb{E}_{S\sim\mathcal{D}}[P_S])\mathbf{1} = 0$$

$$\Rightarrow \quad \mathbb{E}_{S\sim\mathcal{D}}[\psi_{S,J}(w)] = 0$$

$$\min_{w\in\mathbf{R}^d} \mathbb{E}_{S\sim\mathcal{D}}\left[f_{S,J}(w) := f_S(w) + \psi_{S,J}(w)\right], \qquad (\textbf{CSP})$$

Stoch grad: $\nabla f_{S,J}(w^t) := \frac{1}{n}J - \frac{\theta}{n}(J - DF(w^t))P_S\mathbf{1}$

Apply SGD to solve **CSP**?

Controlled Stochastic Reformulation

$$\nabla f_{S,J}(w) \; := \; \frac{1}{n} J - \frac{\theta}{n} \left(J - DF(w) \right) P_S \mathbf{1}$$

$$\begin{aligned}
&\textbf{for } t = 1, \cdots, T \\
&\qquad \textbf{sample } S \sim \mathcal{D} \\
&\qquad J^{t+1} = \textbf{update}(J^t) \\
&\qquad w^{t+1} = w^t - \alpha \nabla f_{S,J^t}(w^t) \\
&\textbf{end for}
\end{aligned}$$

Controlled Stochastic Reformulation

$$\nabla f_{S,J}(w) \; := \; \frac{1}{n} J - \frac{\theta}{n} \left(J - DF(w) \right) P_S \mathbf{1}$$

for $t = 1, \cdots, T$

sample $S \sim \mathcal{D}$

$J^{t+1} = \textbf{update}(J^t)$

$w^{t+1} = w^t - \alpha \nabla f_{S,J^t}(w^t)$

end for

Update to decrease variance

Controlled Stochastic Reformulation

$$\nabla f_{S,J}(w) \; := \; \frac{1}{n} J - \frac{\theta}{n} \left(J - DF(w)\right) P_S \mathbf{1}$$

Update to decrease variance

$$
\begin{aligned}
&\textbf{for } t = 1, \cdots, T \\
&\qquad \textbf{sample } S \sim \mathcal{D} \\
&\qquad J^{t+1} = \textbf{update}(J^t) \\
&\qquad w^{t+1} = w^t - \alpha \nabla f_{S,J^t}(w^t) \\
&\textbf{end for}
\end{aligned}
$$

SGD step

Controlled Stochastic Reformulation

$$\nabla f_{S,J}(w) \; := \; \frac{1}{n} J - \frac{\theta}{n} \left(J - DF(w)\right) P_S \mathbf{1}$$

$$
\begin{aligned}
&\textbf{for } t = 1, \cdots, T \\
&\qquad \textbf{sample } S \sim \mathcal{D} \\
&\qquad J^{t+1} = \textbf{update}(J^t) \\
&\qquad w^{t+1} = w^t - \alpha \nabla f_{S,J^t}(w^t) \\
&\textbf{end for}
\end{aligned}
$$

Update to decrease variance

SGD step

We want:

Unbiased

$$\mathbb{E}[\nabla f_{S,J^t}(w^t)] = \nabla f(w^t)$$

Controlled Stochastic Reformulation

$$\nabla f_{S,J}(w) \; := \; \frac{1}{n} J - \frac{\theta}{n} \left(J - DF(w)\right) P_S \mathbf{1}$$

$$
\begin{aligned}
&\textbf{for } t = 1, \cdots, T \\
&\qquad \textbf{sample } S \sim \mathcal{D} \\
&\qquad J^{t+1} = \textbf{update}(J^t) \\
&\qquad w^{t+1} = w^t - \alpha \nabla f_{S,J^t}(w^t) \\
&\textbf{end for}
\end{aligned}
$$

Update to decrease variance

SGD step

We want:

Unbiased

$$\mathbb{E}[\nabla f_{S,J^t}(w^t)] = \nabla f(w^t)$$

Converges in *L2*

$$\mathbb{E}\left[||\nabla f_{S,J^t}(w^t) - \nabla f(w^t)||_2^2 \right] \longrightarrow 0$$

?

Updating Stoch. Reformulation

How to choose J ?

Idea: Minimize variance

$$
\begin{aligned}
\mathbb{E}||\nabla f_{S,J}(w) - \nabla f(w)||_2^2 &= \mathbb{E}||\frac{1}{n}J(I-\eta P_S)\mathbf{1} - \frac{1}{n}DF(w)(I-\eta P_S)\mathbf{1}||_2^2 \\
&= \frac{1}{n^2}\mathbb{E}||(J-DF(w))(I-\eta P_S)\mathbf{1}||_2^2 \\
&= \frac{1}{n^2}\mathbf{Tr}\left((J-DF(w))^\top (J-DF(w))G\right). \\
&= \frac{1}{n^2}||J-DF(w)||_G^2
\end{aligned}
$$

Where $\quad G := \mathbb{E}\left[(I-\eta P_S)\mathbf{1}\mathbf{1}^\top(I-\eta P_S^\top)\right] \succ 0$

Updating Stoch. Reformulation

How to choose J ?

Idea: Minimize variance

$$
\begin{aligned}
\mathbb{E}||\nabla f_{S,J}(w) - \nabla f(w)||_2^2 &= \mathbb{E}||\frac{1}{n}J(I-\eta P_S)\mathbf{1} - \frac{1}{n}DF(w)(I-\eta P_S)\mathbf{1}||_2^2 \\
&= \frac{1}{n^2}\mathbb{E}||(J-DF(w))(I-\eta P_S)\mathbf{1}||_2^2 \\
&= \frac{1}{n^2}\mathbf{Tr}\left((J-DF(w))^\top (J-DF(w))G\right). \\
&= \frac{1}{n^2}||J-DF(w)||_G^2
\end{aligned}
$$

Where $\quad G := \mathbb{E}\left[(I-\eta P_S)\mathbf{1}\mathbf{1}^\top (I-\eta P_S^\top)\right] \succ 0$

$$
\arg\min_{J\in\mathbb{R}^{d\times n}} \mathbb{E}||\nabla f_{S,J}(w) - \nabla f(w)||_2^2 \;=\; \arg\min_{J\in\mathbb{R}^{d\times n}} ||J-DF(w)||^2
$$

Updating Stoch. Reformulation

How to choose J ?

Idea: Minimize variance

$$\begin{aligned}
\mathbb{E}||\nabla f_{S,J}(w) - \nabla f(w)||_2^2 &= \mathbb{E}||\frac{1}{n}J(I-\eta P_S)\mathbf{1} - \frac{1}{n}DF(w)(I-\eta P_S)\mathbf{1}||_2^2 \\
&= \frac{1}{n^2}\mathbb{E}||(J-DF(w))(I-\eta P_S)\mathbf{1}||_2^2 \\
&= \frac{1}{n^2}\mathbf{Tr}\left((J-DF(w))^\top (J-DF(w))G\right). \\
&= \frac{1}{n^2}||J-DF(w)||_G^2
\end{aligned}$$

Where $\quad G := \mathbb{E}\left[(I-\eta P_S)\mathbf{1}\mathbf{1}^\top (I-\eta P_S^\top)\right] \succ 0$

$$\arg\min_{J\in\mathbb{R}^{d\times n}} \mathbb{E}||\nabla f_{S,J}(w) - \nabla f(w)||_2^2 = \arg\min_{J\in\mathbb{R}^{d\times n}} ||J-DF(w)||^2$$

New idea: Gradually minimize variance

Random Projection of Jacobian

$$J^{t+1} = \arg \min_{J \in \mathbb{R}^{d\times n}, Y \in \mathbb{R}^{d\times \tau}} ||J - DF(w^t)||^2$$
$$\text{subject to } J = J^t + YS^\top$$

Solution:

$$J^{t+1} = J^t - (J^t - DF(w^t))S(S^\top S)^{-1}S^\top$$

This is an equivalent dual viewpoint of "Sketching and projecting the Jacobian"

RMG, P. Richtarik (2015), SIAM. J. Matrix Anal. & Appl.
Randomized Iterative Methods for Linear Systems

A Jacobian Based Method

JacSketch Algorithm

choose distribution $\mathcal{D}$ and the bias-correcting variable $\theta > 0$
choose $\alpha > 0, w^1 \in \mathbb{R}^d, J^1 \in \mathbb{R}^{d \times n}$
for $t = 1, \ldots, T$
  sample $S \sim \mathcal{D}$
  calculate sketch $DF(w^t)S$
  update $J^{t+1} = J^t - (J^t - DF(w^t))S(S^\top S)^{-1} S^\top$
  calculate $g^t = \frac{\theta}{n} J^{t+1}\mathbf{1} + \frac{1-\theta}{n} J^t \mathbf{1}$
  step $w^{t+1} = w^t - \alpha g^t$

Exactly the same method we deduced previously

Proving Convergence

Assumptions for Convergence

Strong Convexity

$$f(w) \geq f(y) + \langle \nabla f(y), w - y \rangle + \frac{\mu}{2} ||w - y||_2^2$$

Smoothness around optimum $\forall C \subset \{1, \ldots, n\} \, \exists \, L_C \geq 0,$

$$||\nabla f_C(w) - \nabla f_C(w^*)||_2^2 \leq 2L_C \left(f_C(w) - f_C(w^*) - \langle \nabla f_C(w^*), w - w^* \rangle \right)$$

$$f_C(w) := \frac{1}{|C|} \sum_{i \in C} f_i(w)$$

Assumptions for Convergence

Strong Convexity

$$f(w) \geq f(y) + \langle \nabla f(y), w - y \rangle + \frac{\mu}{2} ||w - y||_2^2$$

Smoothness around optimum $\forall C \subset \{1, \ldots, n\} \, \exists \, L_C \geq 0,$

$$||\nabla f_C(w) - \nabla f_C(w^*)||_2^2 \leq 2 L_C \left(f_C(w) - f_C(w^*) - \langle \nabla f_C(w^*), w - w^* \rangle \right)$$

$$L_{\max} := \max_{i=1 \ldots n} L_i$$

$$L := L_{\{1, \ldots, n\}}$$

$$f_C(w) := \frac{1}{|C|} \sum_{i \in C} f_i(w)$$

Constants of Convergence I

Expected Smoothness. Let $\mathcal{L}$ be such that

$$\mathbb{E}[||\nabla f_{S,J}(x) - \nabla f_{S,J}(x^*)||_2^2] \;\leq\; \mathcal{L}\,(f(x) - f(x^*)), \quad \forall x$$

Example

- $\mathbb{P}[S = e_i] = 1/n$ for $i = 1, \ldots, n \;\Rightarrow\; \mathcal{L} \;=\; L_{\max}$

Constants of Convergence I

Expected Smoothness. Let $\mathcal{L}$ be such that

$$\mathbb{E}[||\nabla f_{S,J}(x) - \nabla f_{S,J}(x^*)||_2^2] \leq \mathcal{L}\left(f(x) - f(x^*)\right), \quad \forall x$$

Example

- $\mathbb{P}[S = e_i] = 1/n$ for $i = 1, \ldots, n \quad \Rightarrow \mathcal{L} = L_{\max}$
- $\mathbb{P}[S \text{ is invertible}] = 1 \quad \Rightarrow \mathcal{L} = L$

Constants of Convergence I

Expected Smoothness. Let $\mathcal{L}$ be such that

$$\mathbb{E}[||\nabla f_{S,J}(x) - \nabla f_{S,J}(x^*)||_2^2] \;\leq\; \mathcal{L}\,(f(x) - f(x^*)), \quad \forall x$$

Example

- $\mathbb{P}[S = e_i] = 1/n$ for $i = 1, \ldots, n \;\Rightarrow\; \mathcal{L} \;=\; L_{\max}$
- $\mathbb{P}[S \text{ is invertible}] = 1 \;\Rightarrow\; \mathcal{L} \;=\; L$
- $\mathbb{P}[S = I_C] = 1/\binom{n}{\tau}$, for all $C \subset \{1, \ldots, n\}$ with $|C| = \tau$

$$\Rightarrow \quad \mathcal{L} \;=\; \frac{1}{\binom{n-1}{\tau-1}} \max_{i=1,\ldots,n} \sum_{\substack{C \subset \{1,\ldots,n\}, \\ |C| = \tau,\, i \in C}} L_C$$

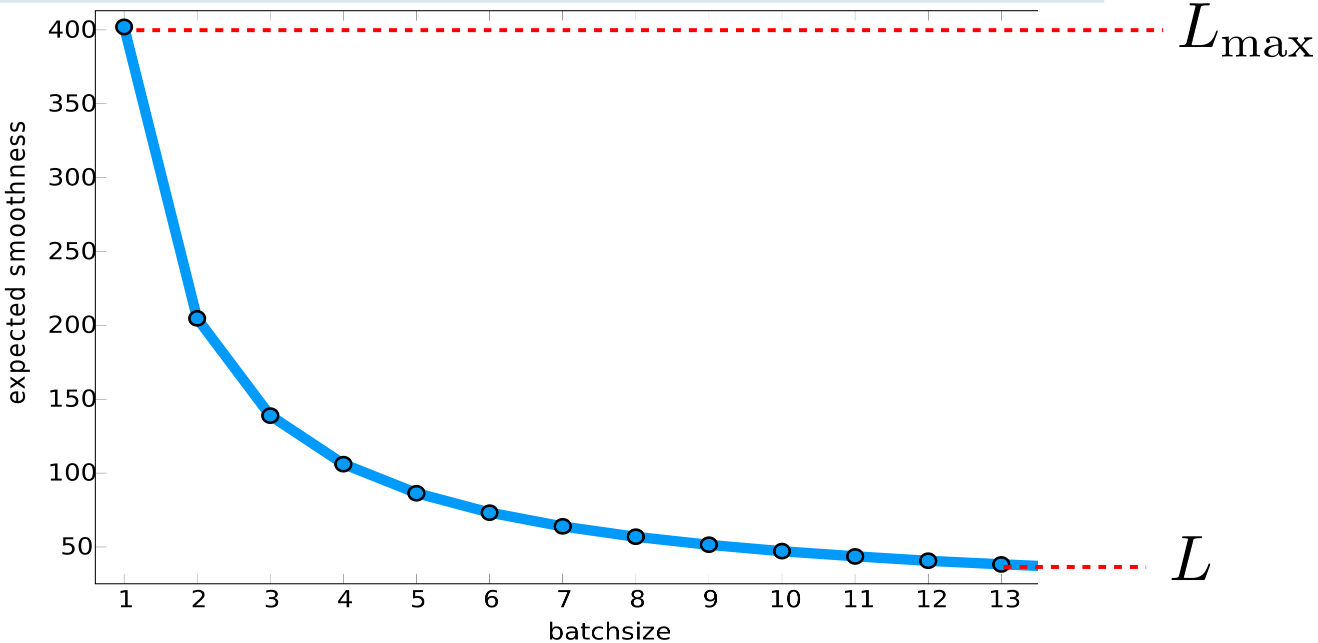
Constants of Convergence I

Expected Smoothness. Let $\mathcal{L}$ be such that

$$\mathbb{E}[||\nabla f_{S,J}(x) - \nabla f_{S,J}(x^*)||_2^2] \;\leq\; \mathcal{L}\,(f(x) - f(x^*)), \quad \forall x$$

$$\mathcal{L} \;=\; \frac{1}{\binom{n-1}{\tau-1}} \max_{i=1,\ldots,n} \sum_{\substack{C \subset \{1,\ldots,n\},\\ |C|=\tau,\; i \in C}} L_C$$

Embodies the trade-off: Cheaper iterates vs less smooth approx.



Constants of Convergence II

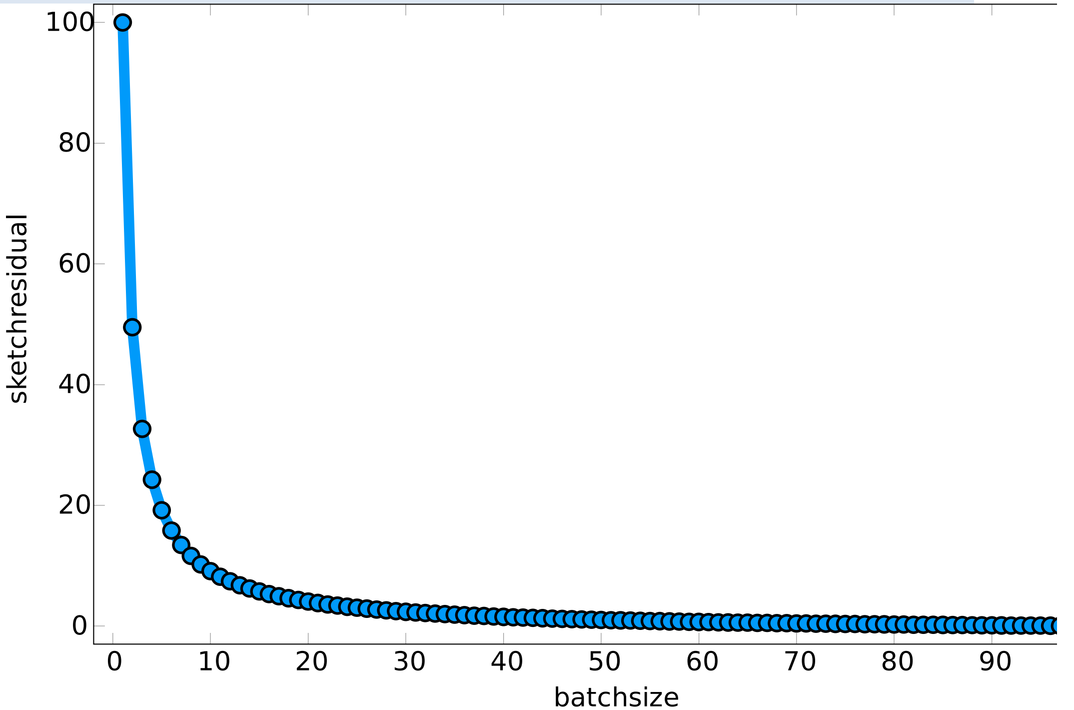
Sketch Residual

$$\rho \quad := \quad \lambda_{\max}\left(\theta^2 \mathbb{E}[P_S \mathbf{1}\mathbf{1}^\top P_S] - \mathbf{1}\mathbf{1}^\top\right)$$

Example

- $\mathbb{P}[S = I_C] = 1/\binom{n}{\tau}$, for all $C \subset \{1,\ldots,n\}$ with $|C| = \tau$

$$\Rightarrow \quad \rho = \frac{n}{\tau}\frac{n-\tau}{n-1}$$



Resumes how much information is lost by the sketch. Does not depend on the data.

Complexity Theorem

Theorem

Let $$\Psi_t := ||w^t - w^*||_2^2 + \sigma ||J^t - DF(w^*)||^2$$

If $(w^t,\, g^t,\, J^t)$ is calculated using the JacSketch algorithm and

$$\alpha \leq \min\left\{\frac{1}{4\mathcal{L}}, \frac{\lambda_{\min}(\mathbb{E}[P_S])}{4\rho\, L_{\max}\lambda_{\max}(\mathbb{E}[P_S])/\, n + \mu}\right\}$$

Complexity Theorem

Theorem

Let $$\Psi_t := ||w^t - w^*||_2^2 + \sigma ||J^t - DF(w^*)||^2$$

If $(w^t,\, g^t,\, J^t)$ is calculated using the JacSketch algorithm and

$$\alpha \leq \min\left\{ \frac{1}{4\mathcal{L}}, \frac{\lambda_{\min}(\mathbb{E}[P_S])}{4\rho\, L_{\max}\lambda_{\max}(\mathbb{E}[P_S])/\, n + \mu} \right\}$$

then $$\mathbb{E}[\Psi_t(\sigma)] \leq (1 - \mu\alpha)\mathbb{E}[\Psi_{t-1}(\sigma)]$$

The resulting iteration complexity is given by

$$t \geq \max\left\{ \frac{4\mathcal{L}}{\mu},\; \frac{1}{\lambda_{\min}(\mathbf{E}[P_S])} + \frac{4\rho L_{\max}}{\mu n} \frac{\lambda_{\max}(\mathbf{E}[P_S])}{\lambda_{\min}(\mathbf{E}[P_S])} \right\} \log\left(\frac{1}{\epsilon}\right)$$

Complexity example I

Corollary (Gradient descent)

If S be invertible with probability one then

$$\begin{aligned} \mathcal{L} &= L \\ \rho &= 0 \\ \mathbb{E}[P_S] &= I \end{aligned}$$

Consequently Theorem gives iteration complexity of

$$t \geq \max\left\{\frac{4L}{\mu},\ 1\right\}\log\left(\frac{1}{\epsilon}\right) = \frac{4L}{\mu}$$

Recovers the classic μ/L convergence rate of Gradient Descent!

Complexity example II

Corollary (Minibatch Saga)

Let $\mathbb{P}[S = I_C] = 1/\binom{n}{\tau}$ for all $C \subset \{1,\ldots,n\}$ with $|C| = \tau$

$$\begin{aligned}
\mathcal{L} &= \frac{1}{\binom{n-1}{\tau-1}} \max_{i=1,\ldots,n} \sum_{\substack{C \subset \{1,\ldots,n\},\\ |C|=\tau,\, i \in C}} L_C \\
\rho &= \frac{n}{\tau}\frac{n-\tau}{n-1} \\
\mathbb{E}[P_S] &= \frac{\tau}{n} I
\end{aligned}$$

Consequently Theorem gives iteration complexity of

$$t \geq \max\left\{\frac{4\mathcal{L}}{\mu}, \; \frac{n}{\tau} + \frac{n-\tau}{(n-1)\tau}\frac{4L_{\max}}{\mu}\right\} \log\left(\frac{1}{\epsilon}\right)$$

First clear speed-up when increasing minibatch size

Complexity example II

Corollary (Minibatch Saga)

Let $\mathbb{P}[S = I_C] = 1/\binom{n}{\tau}$ for all $C \subset \{1,\ldots,n\}$ with $|C| = \tau$

$\mathcal{L}$ $, n\},$ L_C

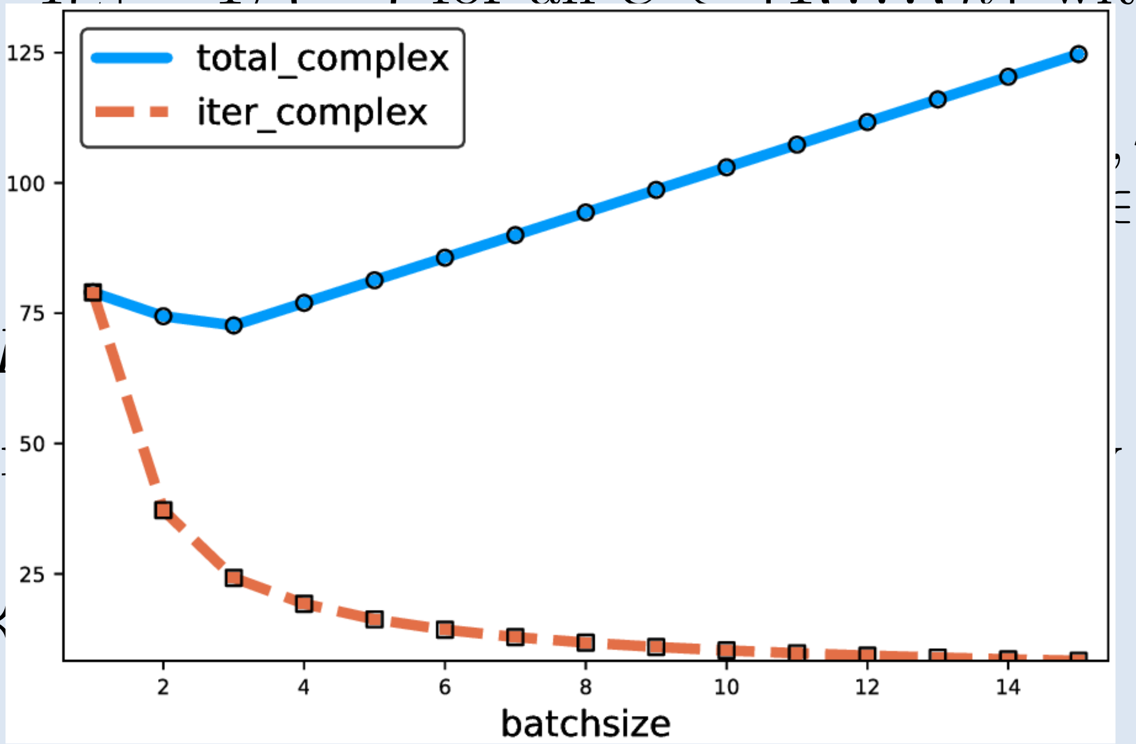
$\in C$

ρ

$\mathbb{E}[$

Conseque of

$t \geq \max$



First clear speed-up in complexity when increasing minibatch size

Complexity example II

Corollary (Minibatch Saga)

Let $\mathbb{P}[S = I_C] = 1/\binom{n}{\tau}$ for all $C \subset \{1,\ldots,n\}$ with $|C| = \tau$

$\mathcal{L}$ $, n\},$ L_C

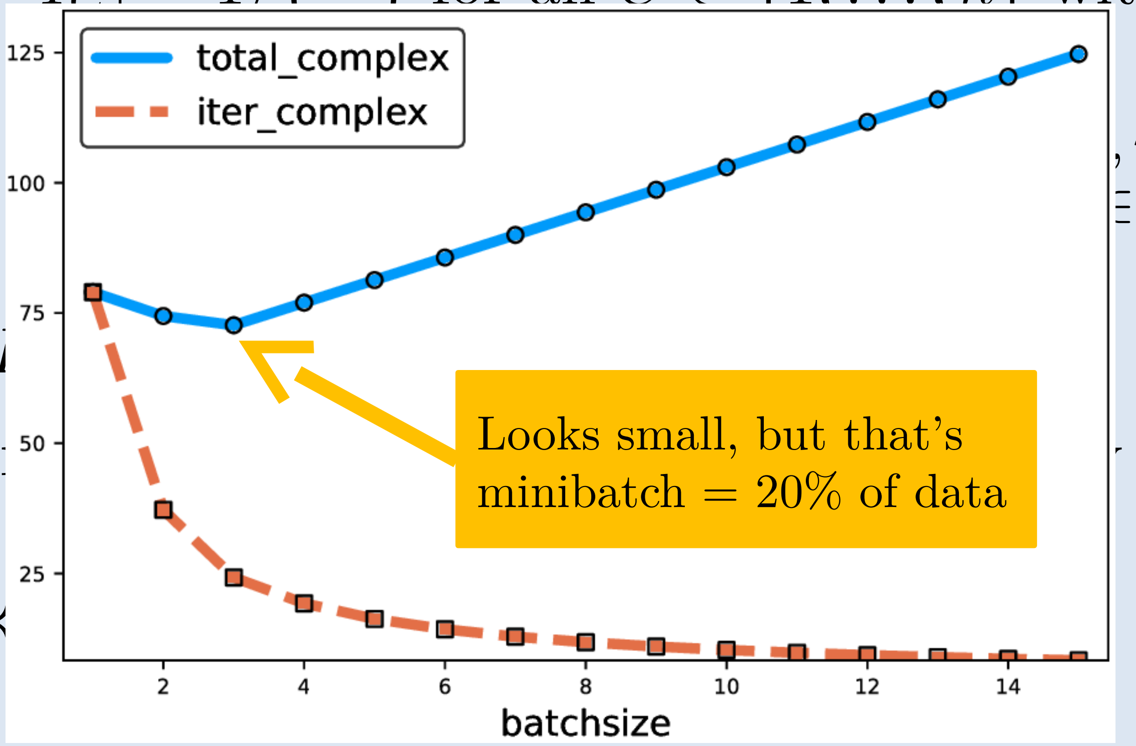
$\in C$

ρ

$\mathbb{E}[$

Conseque… of

$t \geq \max$



First clear speed-up in complexity when increasing minibatch size

Further/Future results

Non-uniform samplings

Optimal probabilities for sampling with

$$O\left(n + \frac{\sum_{i=1} L_i}{n\mu}\right)$$

✓

Inner-outer loop SVRG type methods

✓

Sparse Johnson-Lindenstraus sketches

?

For minimizing true expectations

?

Schmidt, Babanezhad, Ahmed, Defazio, Clifton, Sarkar. AISTATS, 2015 **Non-Uniform Stochastic Average Gradient Method for Training Conditional Random Fields**.

To be continued....

M. Schmidt, N. Le Roux, F. Bach (2016), Mathematical Programming
Minimizing Finite Sums with the Stochastic Average Gradient.

RMG, P. Richtarik, F. Bach (2018), preprint online
Stochastic quasi-gradient methods: Variance reduction via Jacobian sketching

T. Hofmann, A. Lucchi, S. Lacoste-Julien, B. McWilliams (2016), NIPS,
Variance Reduced Stochastic Gradient Descent with Neighbors