Statistical inference with stochastic gradient descent

Panagiotis (Panos) Toulis
panos.toulis@chicagobooth.edu

Econometrics and Statistics
University of Chicago, Booth School of Business
Optimization and estimation are complementary.

amazing new Yahoo dataset for machine learning research: 1.5 TB of Yahoo news, plus 20M user interactions
webscope.sandbox.yahoo.com/catalog.php?da...
Optimization and estimation are complementary.

A unique method that crosses the optimization-statistics boundary:

**stochastic gradient descent (SGD)**

Special case of **stochastic approximation** (Robbins & Monro, 1951).
Statistical estimation gave a new form of optimization problems:

$$\max_{\theta} L(\theta) \Leftrightarrow \max_{\theta} \sum_{i=1}^{N} l_i(\theta),$$

where $l_i$ is log-likelihood of $\theta$ at $i$th datapoint.
SGD in optimization and statistics

- Statistical estimation gave a new form of optimization problems:
  \[
  \max_{\theta} L(\theta) \iff \max_{\theta} \sum_{i=1}^{N} l_i(\theta),
  \]
  where \( l_i \) is log-likelihood of \( \theta \) at \( i \)th datapoint.

- Classical optimization methods, such as gradient descent,
  \[
  \theta_n = \theta_{n-1} + \gamma_n \nabla L(\theta_{n-1}),
  \]
  fail when computation of gradient is expensive (e.g., \( N \) large).
Statistical estimation gave a new form of optimization problems:

$$\max_{\theta} L(\theta) \iff \max_{\theta} \sum_{i=1}^{N} l_i(\theta),$$

where $l_i$ is log-likelihood of $\theta$ at $i$th datapoint.

Classical optimization methods, such as gradient descent,

$$\theta_n = \theta_{n-1} + \gamma_n \nabla L(\theta_{n-1}),$$

fail when computation of gradient is expensive (e.g., $N$ large).

SGD has emerged as the most versatile optimization method:

$$\theta_n = \theta_{n-1} + \gamma_n \nabla l_{i_n}(\theta_{n-1}),$$

where $i_n$ is uniformly sampled from $[1, 2, \ldots, N]$. 
Statistical estimation gave a new form of optimization problems:

$$\max_{\theta} L(\theta) \iff \max_{\theta} \sum_{i=1}^{N} l_i(\theta),$$

where $l_i$ is log-likelihood of $\theta$ at $i$th datapoint.

Classical optimization methods, such as gradient descent,

$$\theta_n = \theta_{n-1} + \gamma_n \nabla L(\theta_{n-1}),$$

fail when computation of gradient is expensive (e.g., $N$ large).

SGD has emerged as the most versatile optimization method:

$$\theta_n = \theta_{n-1} + \gamma_n \nabla l_{i_n}(\theta_{n-1}),$$

where $i_n$ is uniformly sampled from $[1, 2, \ldots, N]$.

However, SGD is not popular in statistics. Why?
SGD in optimization and statistics

- Statistical estimation gave a new form of optimization problems:

\[
\max_{\theta} L(\theta) \iff \max_{\theta} \sum_{i=1}^{N} l_i(\theta),
\]

where \( l_i \) is log-likelihood of \( \theta \) at \( i \)th datapoint.

- Classical optimization methods, such as gradient descent,

\[
\theta_n = \theta_{n-1} + \gamma_n \nabla L(\theta_{n-1}),
\]

fail when computation of gradient is expensive (e.g., \( N \) large).

- SGD has emerged as the most versatile optimization method:

\[
\theta_n = \theta_{n-1} + \gamma_n \nabla l_{i_n}(\theta_{n-1}),
\]

where \( i_n \) is uniformly sampled from \( [1, 2, \ldots, N] \).

- However, SGD is not popular in statistics. Why?

SGD has not been reliable for statistical estimation.
Motivation: modeling flight ticket booking

- $Y = \#\text{bookings}; \ X = \text{covariates}; \ Y \sim \text{Poisson}(e^{X'\theta^*})$.

- **Goal:** use i.i.d. data $(X_1, Y_1), (X_2, Y_2), \ldots$, to estimate $\theta^*$. 

Example: $\theta_0 = 0, Y_1 = Y_2 = 1001, X_n = 1$.

First iteration: $\theta_1 = 0 + 1 \cdot (1001 - 1) = 1000$.

Second iteration: $\theta_2 = \theta_1 + 1/2 (1001 - e^{1000}) = -\infty$.

Standard SGD procedures are often numerically unstable.
Motivation: modeling flight ticket booking

- $Y =$ #bookings; $X =$ covariates; $Y \sim \text{Poisson}(e^{X'\theta^*})$.
- **Goal:** use i.i.d. data $(X_1, Y_1), (X_2, Y_2), \ldots$, to estimate $\theta^*$.
- Estimation through SGD uses discrepancies (observed $n$ − expected $n$):

$$
\theta_n = \theta_{n-1} + \frac{1}{n} (Y_n - e^{X_n'\theta_{n-1}}) X_n.
$$
Motivation: modeling flight ticket booking

- $Y = \#\text{bookings}; X = \text{covariates}; Y \sim \text{Poisson}(e^{X'\theta*})$.
- **Goal:** use i.i.d. data $(X_1, Y_1), (X_2, Y_2), \ldots$, to estimate $\theta_*$.
- Estimation through SGD uses discrepancies (observed $n - \text{expected } n$):

  $$\theta_n = \theta_{n-1} + \frac{1}{n}(Y_n - e^{X_n'\theta_{n-1}})X_n.$$ 

**Example:** $\theta_0 = 0, Y_1 = Y_2 = 1001, X_n = 1$.

- First iteration:

  $$\theta_1 = 0 + 1 \cdot (1001 - 1) = 1000.$$ 

- Second iteration:

  $$\theta_2 = \theta_1 + \frac{1}{2}(1001 - e^{1000}) = -\infty.$$ 

- Standard SGD procedures are often numerically unstable.
Estimation through SGD with \textbf{implicit} update:

\[ \theta_n = \theta_{n-1} + \frac{1}{n} (Y_n - e^{X_n^t \theta_n}) X_n. \]

(\textbf{See} intuition. \textbf{See} computation.)
Estimation through SGD with **implicit** update:

\[
\theta_n = \theta_{n-1} + \frac{1}{n} (Y_n - e^{X'_n \theta_n}) X_n.
\]

*(See intuition. See computation.)*

**Example:** \(\theta_0 = 0, Y_1 = Y_2 = 1001, X_n = 1\).

- **First iteration,**

  \[
  \theta_1 = 0 + 1 \cdot (1001 - e^{\theta_1}).
  \]

  Thus, \(\theta_1 \approx \log(994) \approx 6.902\).

- **Second iteration,**

  \[
  \theta_2 = 6.902 + \frac{1}{2} (1001 - e^{\theta_2}).
  \]

  Thus, \(\theta_2 \approx \log(1001) \approx 6.909\).
Related work and contributions

- Work in stochastic optimization that has considered implicit updates:
  - normalized least mean squares (Nagumo & Noda, 1967); (Slock, 1993);
  - incremental proximal method (Bertsekas, 2011);
  - stochastic proximal gradient (Duchi & Singer, 2009); (Rosasco, 2014);
  - implicit online learning (Kivinen & Warmuth, 1997); (Kulis & Bartlett, 2010).
Related work and contributions

- Work in stochastic optimization that has considered implicit updates:
  - normalized least mean squares (Nagumo & Noda, 1967); (Slock, 1993);
  - incremental proximal method (Bertsekas, 2011);
  - stochastic proximal gradient (Duchi & Singer, 2009); (Rosasco, 2014);
  - implicit online learning (Kivinen & Warmuth, 1997); (Kulis & Bartlett, 2010).

- Three main contributions of our work:
  1. We analyze statistical efficiency of SGD-based estimators.
  2. We develop theory, methods and code for implicit SGD;
  3. ...towards statistical inference with SGD methods.
Notation and definitions

- \( Y \in \mathbb{R}^m \) outcome, \( X \in \mathbb{R}^p \) covariate, model \( f \), true param. \( \theta_\star \in \mathbb{R}^p \):
  \[
  Y|X \sim f(Y; X, \theta_\star).
  \]

- Fisher information matrix \((p \times p)\):
  \[
  \mathcal{I}(\theta) = -\mathbb{E} \left( \nabla^2 \log f(Y; X, \theta) \right) \succeq 0.
  \]
Notation and definitions

- $Y \in \mathbb{R}^m$ outcome, $X \in \mathbb{R}^p$ covariate, model $f$, true param. $\theta^* \in \mathbb{R}^p$:

  $$Y|X \sim f(Y; X, \theta^*).$$

- Fisher information matrix $(p \times p)$:

  $$\mathcal{I}(\theta) = -\mathbb{E} \left( \nabla^2 \log f(Y; X, \theta) \right) \succeq 0.$$ 

- Quantity $\mathcal{I}(\theta^*)^{-1}$ is the Cramér-Rao bound. Suppose $\hat{\theta}_n \to \theta^*$, then

  $$\lim_{n \to \infty} n \text{Var}(\hat{\theta}_n) = \mathcal{I}(\theta^*)^{-1}$$

  optimal efficiency; 

  $$\mathcal{I}(\theta^*)^{-1}$$

  efficiency loss; 

  $$\infty$$

  inefficiency.

- Eigenvalues $\lambda_j \in \text{eig}(\mathcal{I}(\theta^*))$, $\lambda_{\min} = \min_j \lambda_j$, $\lambda_{\max} = \max_j \lambda_j$. 

Panagiotis (Panos) Toulis

Statistical properties of stochastic gradient methods 7 / 24
We consider a stream of i.i.d. data $(X_i, Y_i) \sim f_{\theta^*}$, $i = 1, 2, \ldots$

Explicit SGD estimator of $\theta^*$ after $n$ datapoints is

$$\theta_{n}^{\text{ex}} = \theta_{n-1}^{\text{ex}} + \gamma_n \nabla \log f(Y_n; X_n, \theta_{n-1}^{\text{ex}}),$$

where $\gamma_n = \gamma_1/n$, where $\gamma_1 > 0$, is the learning rate parameter.
We consider a stream of i.i.d. data \((X_i, Y_i) \sim f_{\theta^*}, i = 1, 2, \ldots\)

**Explicit SGD estimator of** \(\theta^*\) **after** \(n\) **datapoints is**

\[
\theta_{n}^{\text{ex}} = \theta_{n-1}^{\text{ex}} + \gamma_n \nabla \log f(Y_n; X_n, \theta_{n-1}^{\text{ex}}),
\]

where \(\gamma_n = \gamma_1/n\), where \(\gamma_1 > 0\), is the learning rate parameter.

**Implicit SGD estimator:**

\[
\theta_{n}^{\text{im}} = \theta_{n-1}^{\text{im}} + \gamma_n \nabla \log f(Y_n; X_n, \theta_{n}^{\text{im}}). \quad \text{Bayesian interpretation}
\]
We consider a stream of i.i.d. data \((X_i, Y_i) \sim f_{\theta_\star}, i = 1, 2, \ldots\)

**Explicit SGD estimator of \(\theta_\star\) after \(n\) datapoints is**

\[
\theta_{n}^{\text{ex}} = \theta_{n-1}^{\text{ex}} + \gamma_n \nabla \log f(Y_n; X_n, \theta_{n-1}^{\text{ex}}),
\]

where \(\gamma_n = \gamma_1/n\), where \(\gamma_1 > 0\), is the learning rate parameter.

**Implicit SGD estimator:**

\[
\theta_{n}^{\text{im}} = \theta_{n-1}^{\text{im}} + \gamma_n \nabla \log f(Y_n; X_n, \theta_{n}^{\text{im}}).
\]

**Stochastic approximation theory (for explicit procedures only):**

\[
\mathbb{E} (\nabla \log f(Y; X, \theta_{\infty})) = 0 \iff \theta_{\infty} = \theta_\star.
\]
We consider a stream of i.i.d. data \((X_i, Y_i) \sim f_{\theta^*}, i = 1, 2, \ldots\)

**Explicit SGD estimator of** \(\theta^*\) **after** \(n\) **datapoints is**

\[
\theta_{n}^{\text{ex}} = \theta_{n-1}^{\text{ex}} + \gamma_n \nabla \log f(Y_n; X_n, \theta_{n-1}^{\text{ex}}),
\]

where \(\gamma_n = \gamma_1/n\), where \(\gamma_1 > 0\), is the learning rate parameter.

**Implicit SGD estimator:**

\[
\theta_{n}^{\text{im}} = \theta_{n-1}^{\text{im}} + \gamma_n \nabla \log f(Y_n; X_n, \theta_{n}^{\text{im}}).
\]

**Stochastic approximation theory (for explicit procedures only):**

\[
\mathbb{E} (\nabla \log f(Y; X, \theta_\infty)) = 0 \iff \theta_\infty = \theta^*.
\]

**Note:** When \#datapoints is fixed the characteristic equation holds in-sample wrt the empirical distribution of \((X, Y)\); in this case \(\theta^*\) is an empirical minimizer (e.g., MLE or MAP).
Focus on sampling variability

This talk focuses on:
- Asymptotic variance of SGD in closed form; statistical implications.
- Asymptotic normality, especially of implicit SGD thanks to its stability.

possible datasets of size $n$

sampling distribution:
$E|\theta_n-\theta^*|, \text{Var}(\theta_n)$
Focus on sampling variability

This talk focuses on:

- Asymptotic variance of SGD in closed form; statistical implications.
- Asymptotic normality, especially of implicit SGD thanks to its stability.
Numerical stability: illustration on normal model

- Suppose $Y|X \sim \mathcal{N}(X^\top \theta_\star, 1)$. Then, $\nabla \log f(Y; X, \theta) = (Y - X^\top \theta)X$.
- Fisher information: $\mathcal{I}(\theta) = -\mathbb{E} (\nabla^2 \log f(Y; X, \theta)) = \mathbb{E} (XX^\top)$. 
Numerical stability: illustration on normal model

- Suppose $Y|X \sim \mathcal{N}(X^\top \theta_*, 1)$. Then, $\nabla \log f(Y; X, \theta) = (Y - X^\top \theta)X$.
- Fisher information: $\mathcal{I}(\theta) = -\mathbb{E} \left( \nabla^2 \log f(Y; X, \theta) \right) = \mathbb{E} \left( XX^\top \right)$.
- Explicit SGD estimator:
  \[
  \theta_{n}^{\text{ex}} = (\mathbb{I} - \gamma_n X_n X_n^\top) \theta_{n-1}^{\text{ex}} + \gamma_n Y_n X_n.
  \]
Numerical stability: illustration on normal model

- Suppose $Y|X \sim \mathcal{N}(X^\top \theta_*, 1)$. Then, $\nabla \log f(Y; X, \theta) = (Y - X^\top \theta)X$.
- Fisher information: $\mathcal{I}(\theta) = -\mathbb{E} \left( \nabla^2 \log f(Y; X, \theta) \right) = \mathbb{E} (XX^\top)$.

- Explicit SGD estimator:
  \[
  \theta_{n+1}^{\text{ex}} = (\mathbb{I} - \gamma_n X_n X_n^\top) \theta_{n-1}^{\text{ex}} + \gamma_n Y_n X_n.
  \]

- Implicit SGD estimator:
  \[
  \theta_{n+1}^{\text{im}} = \theta_{n-1}^{\text{im}} + \gamma_n (Y_n - X_n^\top \theta_{n-1}^{\text{im}}) X_n \\
  = (\mathbb{I} + \gamma_n X_n X_n^\top)^{-1} \theta_{n-1}^{\text{im}} + \gamma_n (\mathbb{I} + \gamma_n X_n X_n^\top)^{-1} Y_n X_n.
  \]
In explicit SGD, solve recursively for $\theta_{ex}^n$ to derive:

$$
\theta_{ex}^n = P_n^0 \theta_{ex}^0 + \sum_{i=1}^{n} \gamma_i P_{i+1}^n Y_i X_i,
$$

where $P_i^n = (I - \gamma_n X_n X_n^T) \cdots (I - \gamma_i X_i X_i^T)$. 

(1 - \gamma_1 \lambda_j) (1 - \gamma_2 \lambda_j^2) \cdots (1 - \gamma_n \lambda_j^n) \in \text{eig}\left(E(P_1^n)\right)$. 

For stability it is desirable that $|1 - \gamma_1 \lambda_j| < 1 \Rightarrow \gamma_1 < \frac{2}{\lambda_{\text{max}}}$. 

If $\gamma_1 > \frac{2}{\lambda_{\text{max}}}$, then $\max_{n>0} \{\text{eig}(E(P_1^n))\} = O\left(\frac{2 \gamma_1 \lambda_{\text{max}}}{\sqrt{\gamma_1 \lambda_{\text{max}}}}\right)$. 

In explicit SGD, solve recursively for $\theta_{ex}^n$ to derive:

$$\theta_{ex}^n = P^n_1 \theta_{0ex} + \sum_{i=1}^{n} \gamma_i P^n_{i+1} Y_i X_i,$$

where $P^n_i = (I - \gamma_n X_n X_n^T) \cdots (I - \gamma_i X_i X_i^T)$.

- $P^n_1$ discounts initial conditions $\theta_{0ex}$. Its “size” is critical for stability:

  $$(1 - \gamma_1 \lambda_j)(1 - \frac{\gamma_1}{2} \lambda_j) \cdots (1 - \frac{\gamma_1}{n} \lambda_j) \in \text{eig}(E(P^n_1)).$$
In explicit SGD, solve recursively for $\theta_n^{ex}$ to derive:

$$\theta_n^{ex} = P_1^n \theta_0^{ex} + \sum_{i=1}^{n} \gamma_i P_{i+1}^{n} Y_i X_i,$$

where $P_i^n = (I - \gamma_n X_n X_n^T) \cdots (I - \gamma_i X_i X_i^T)$.

$P_1^n$ discounts initial conditions $\theta_0^{ex}$. Its “size” is critical for stability:

$$(1 - \gamma_1 \lambda_j)(1 - \frac{\gamma_1}{2} \lambda_j) \cdots (1 - \frac{\gamma_1}{n} \lambda_j) \in \text{eig}(E(P_1^n)).$$

For stability it is desirable that

$$|1 - \gamma_1 \lambda_j| < 1 \Rightarrow \gamma_1 < 2/\lambda_{\text{max}}.$$

If $\gamma_1 > 2/\lambda_{\text{max}}$, then

$$\max_{n>0} \max \{ \text{eig}(E(P_1^n)) \} = O(2^{\gamma_1 \lambda_{\text{max}}}/\sqrt{\gamma_1 \lambda_{\text{max}}}).$$
In implicit SGD, solve recursively for $\theta^\text{im}_n$ to derive:

$$
\theta^\text{im}_n = Q^n_1 \theta^\text{im}_0 + \sum_{i=1}^{n} \gamma_i Q^n_{i+1} Q^i Y_i X_i.,
$$

where $Q^n_i = (I + \gamma_n X_n X_n^\top)^{-1} \cdots (I + \gamma_i X_i X_i^\top)^{-1}$. 

Discounts initial $\theta^\text{im}_0$. Its “size” is critical for stability:

$$
(1 + \gamma_1 \lambda_j)^{-1} \cdots (1 + \gamma_n \lambda_j)^{-1} \in \text{eig}(E(Q^n_1)).
$$

Unconditionally stable! For any $\gamma_1 > 0$:

$$
\max_{n>0} \max_{\lambda_j} \{\text{eig}(E(Q^n_1))\} = O(1).
$$

Illustrate with Simulation.
In implicit SGD, solve recursively for $\theta^\text{im}_n$ to derive:

$$\theta^\text{im}_n = Q_1^n \theta^\text{im}_0 + \sum_{i=1}^{n} \gamma_i Q_i^n Q_i^i Y_i X_i,$$

where $Q_i^n = (\mathbb{I} + \gamma_n X_n X_n^\top)^{-1} \cdots (\mathbb{I} + \gamma_i X_i X_i^\top)^{-1}$.

$Q_1^n$ discounts initial $\theta^\text{im}_0$. Its “size” is critical for stability:

$$(1 + \gamma_1 \lambda_j)^{-1} (1 + \frac{\gamma_1}{2} \lambda_j)^{-1} \cdots (1 + \frac{\gamma_1}{n} \lambda_j)^{-1} \in \text{eig}(\mathbb{E}(Q_1^n)),$$

Unconditionally stable! For any $\gamma_1 > 0$, $\max_{n > 0} \text{max}_{\text{eig}(\mathbb{E}(Q_1^n))} = O(1)$.
In implicit SGD, solve recursively for $\theta_{m}^{\text{im}}$ to derive:

$$
\theta_{n}^{\text{im}} = Q_{1}^{n} \theta_{0}^{\text{im}} + \sum_{i=1}^{n} \gamma_{i} Q_{i+1}^{n} Q_{i}^{i} Y_{i} X_{i} \cdot ,
$$

where $Q_{i}^{n} = (I + \gamma_{n} X_{n} X_{n}^{\top})^{-1} \cdots (I + \gamma_{i} X_{i} X_{i}^{\top})^{-1}$.

$Q_{1}^{n}$ discounts initial $\theta_{0}^{\text{im}}$. Its “size” is critical for stability:

$$
(1 + \gamma_{1} \lambda_{j})^{-1} (1 + \frac{\gamma_{1}}{2} \lambda_{j})^{-1} \cdots (1 + \frac{\gamma_{1}}{n} \lambda_{j})^{-1} \in \text{eig}(\mathbb{E}(Q_{1}^{n})).
$$

Unconditionally stable! For any $\gamma_{1} > 0$

$$
\max_{n>0} \max \{ \text{eig}(\mathbb{E}(Q_{1}^{n})) \} = O(1).
$$

Illustrate with Simulation.
Suppose $2\gamma_1 I(\theta^\star) - I > 0$. Then, the variance of implicit SGD satisfies
\[ n \bar{\text{Var}}(\theta_n^{\text{im}}) \to \gamma_1^2 (2\gamma_1 I(\theta^\star) - I)^{-1} I(\theta^\star). \]

The explicit SGD estimator has the same asymptotic efficiency.
Theorem (Toulis et. al., 2014

Suppose $2\gamma_1 \mathcal{I}(\theta^\star) - \mathbb{I} > 0$. Then, the variance of implicit SGD satisfies

$$n \text{Var}(\theta_{im}^n) \to \gamma_1^2 (2\gamma_1 \mathcal{I}(\theta^\star) - \mathbb{I})^{-1} \mathcal{I}(\theta^\star).$$

The explicit SGD estimator has the same asymptotic efficiency.

- Condition $2\gamma_1 \mathcal{I}(\theta^\star) - \mathbb{I} > 0$ implies the requirement

$$\gamma_1 > 1/(2\lambda_{\text{min}}).$$

- If $\gamma_1 < 1/(2\lambda_{\text{min}})$ then arbitrary inefficiency, e.g., $n^c \text{Var}(\theta_{im}^n) \to \infty$. 

- Generalization & Nonasymptotics
Summarizing the constraints

<table>
<thead>
<tr>
<th></th>
<th>Explicit SGD</th>
<th>Implicit SGD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stability</td>
<td>( \gamma_1 &lt; \frac{2}{\lambda_{\text{max}}} )</td>
<td>-</td>
</tr>
<tr>
<td>Efficiency</td>
<td>( \gamma_1 &gt; \frac{1}{(2\lambda_{\text{min}})} )</td>
<td>( \gamma_1 &gt; \frac{1}{(2\lambda_{\text{min}})} )</td>
</tr>
</tbody>
</table>
Summarizing the constraints

<table>
<thead>
<tr>
<th></th>
<th>Explicit SGD</th>
<th>Implicit SGD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stability</td>
<td>$\gamma_1 &lt; 2/\lambda_{\text{max}}$</td>
<td>-</td>
</tr>
<tr>
<td>Efficiency</td>
<td>$\gamma_1 &gt; 1/(2\lambda_{\text{min}})$</td>
<td>$\gamma_1 &gt; 1/(2\lambda_{\text{min}})$</td>
</tr>
</tbody>
</table>

- The requirements for explicit SGD are hard to reconcile.
  - e.g., $\mathbb{E} \left( \frac{\lambda_{\text{max}}}{\lambda_{\text{min}}} \right) = O(\log p)$ in $p \times p$ std normal random matrix (Edelman, 1988)
Efficiency loss of SGD

- Let \( \Sigma_{\theta^*,\gamma_1} \triangleq \lim n \Var(\theta_n^{im}) = \gamma_1^2 (2\gamma_1 I(\theta^*) - \mathbb{I})^{-1} I(\theta^*). \)
Efficiency loss of SGD

- Let $\Sigma_{\theta_*,\gamma_1} \triangleq \lim_n n \nabla \text{Var}(\theta_n) = \gamma_1^2 (2\gamma_1 \mathcal{I}(\theta_*) - \mathbb{I})^{-1} \mathcal{I}(\theta_*)$.
- It follows,

$$\frac{\gamma_1^2 \lambda_j}{2\gamma_1 \lambda_j - 1} \in \text{eig}(\Sigma_{\theta_*,\gamma_1}), \text{ where } \lambda_j \in \text{eig}(\mathcal{I}(\theta_*)).$$
Efficiency loss of SGD

- Let $\Sigma_{\theta^*,\gamma_1} \triangleq \lim n \text{Var}(\theta_n^{\text{im}}) = \gamma_1^2 (2\gamma_1 \mathcal{I}(\theta^*) - I)^{-1} \mathcal{I}(\theta^*)$.
- It follows,

$$
\frac{\gamma_1^2 \lambda_j}{2\gamma_1 \lambda_j - 1} \in \text{eig}(\Sigma_{\theta^*,\gamma_1}), \text{ where } \lambda_j \in \text{eig}(\mathcal{I}(\theta^*)).
$$

- Implies efficiency loss because $\mathcal{I}(\theta^*)^{-1}$ is optimal (Cramér-Rao bound).

$$
\frac{\gamma_1^2 \lambda_j}{2\gamma_1 \lambda_j - 1} \geq \frac{1}{\lambda_j} \Rightarrow \Sigma_{\theta^*,\gamma_1} \succeq \mathcal{I}(\theta^*)^{-1}.
$$
Efficiency loss of SGD

- Let $\Sigma_{\theta^*,\gamma_1} \triangleq \lim n \text{Var}(\theta_n) = \gamma_1^2(2\gamma_1 \mathcal{I}(\theta^*) - I)^{-1}\mathcal{I}(\theta^*)$.
- It follows,

$$\frac{\gamma_1^2 \lambda_j}{2\gamma_1 \lambda_j - 1} \in \text{eig}(\Sigma_{\theta^*,\gamma_1}), \text{ where } \lambda_j \in \text{eig}(\mathcal{I}(\theta^*)).$$

- Implies efficiency loss because $\mathcal{I}(\theta^*)^{-1}$ is optimal (Cramér-Rao bound).
- No efficiency loss only when $\lambda_j = \lambda$ and $\gamma_1 = 1/\lambda$. 

In practice, large efficiency loss because $\lambda_{\text{max}} \gg \lambda_{\text{min}}$ (spectral gap). 

Panagiotis (Panos) Toulis

Statistical properties of stochastic gradient methods 15 / 24
Efficiency loss of SGD

- Let $\Sigma_{\theta^*, \gamma_1} \triangleq \lim n \text{Var}(\theta_n^{\text{im}}) = \gamma_1^2 (2\gamma_1 I(\theta^*) - I)^{-1} I(\theta^*)$.
- It follows,

$$\frac{\gamma_1^2 \lambda_j}{2\gamma_1 \lambda_j - 1} \in \text{eig}(\Sigma_{\theta^*, \gamma_1}), \text{ where } \lambda_j \in \text{eig}(I(\theta^*)).$$

$$\frac{\gamma_1^2 \lambda_j}{2\gamma_1 \lambda_j - 1} \geq \frac{1}{\lambda_j} \Rightarrow \Sigma_{\theta^*, \gamma_1} \succeq I(\theta^*)^{-1}.$$

- Implies efficiency loss because $I(\theta^*)^{-1}$ is optimal (Cramér-Rao bound).
- No efficiency loss only when $\lambda_j = \lambda$ and $\gamma_1 = 1/\lambda$.
- In practice, large efficiency loss because $\lambda_{\text{max}} \gg \lambda_{\text{min}}$ (spectral gap).
Another principled way to set the optimal rate:

\[ \gamma_1^* = \arg \min_{\gamma_1} \text{tr}(\Sigma_{\theta^*}, \gamma_1) \iff \gamma_1^* = \arg \min_{\gamma_1} \sum_{j=1}^{p} \frac{\gamma_1^2 \lambda_j}{2\gamma_1 \lambda_j - 1}. \]
Another principled way to set the optimal rate:

\[ \gamma_1^* = \arg \min_{\gamma_1} \text{tr}(\Sigma_{\theta^*, \gamma_1}) \iff \gamma_1^* = \arg \min_{\gamma_1} \sum_{j=1}^{p} \frac{\gamma_1^2 \lambda_j}{2\gamma_1 \lambda_j - 1}. \]

Note \( \frac{\gamma_1^2 \lambda_j}{2\gamma_1 \lambda_j - 1} \approx \frac{\gamma_1}{2} \) if \( 2\gamma_1 \lambda_j >> 1 \).
Optimal rates

• Another principled way to set the optimal rate:

\[ \gamma_1^* = \arg\min_{\gamma_1} \text{tr}(\Sigma_{\theta_*}, \gamma_1) \Leftrightarrow \gamma_1^* = \arg\min_{\gamma_1} \sum_{j=1}^{p} \frac{\gamma_1^2 \lambda_j}{2\gamma_1 \lambda_j - 1}. \]

• Note \( \frac{\gamma_1^2 \lambda_j}{2\gamma_1 \lambda_j - 1} \approx \frac{\gamma_1}{2} \) if \( 2\gamma_1 \lambda_j >> 1 \).

• If \( \gamma_1 > 1/(2\lambda_{\min}) \),

\[ \text{tr}(\Sigma_{\theta_*}, \gamma_1) \approx (p - 1)\frac{\gamma_1}{2} + \frac{\gamma_1^2 \lambda_{\min}}{2\gamma_1 \lambda_{\min} - 1}. \]
Another principled way to set the optimal rate:

\[ \gamma_1^* = \arg \min_{\gamma_1} \text{tr}(\Sigma_{\theta^*}, \gamma_1) \Leftrightarrow \gamma_1^* = \arg \min_{\gamma_1} \sum_{j=1}^{p} \frac{\gamma_1^2 \lambda_j}{2 \gamma_1 \lambda_j - 1}. \]

Note \( \frac{\gamma_1^2 \lambda_j}{2 \gamma_1 \lambda_j - 1} \approx \frac{\gamma_1}{2} \) if \( 2 \gamma_1 \lambda_j \gg 1 \).

If \( \gamma_1 > 1/(2 \lambda_{\text{min}}) \),

\[ \text{tr}(\Sigma_{\theta^*}, \gamma_1) \approx (p - 1) \frac{\gamma_1}{2} + \frac{\gamma_1^2 \lambda_{\text{min}}}{2 \gamma_1 \lambda_{\text{min}} - 1}. \]

If \( \gamma_1 \gg 1/(2 \lambda_{\text{min}}) \),

\[ \text{tr}(\Sigma_{\theta^*}, \gamma_1) \approx p \frac{\gamma_1}{2} \quad (\text{In fact, } \Sigma_{\theta^*}, \gamma_1 \approx \frac{\gamma_1}{2} I). \]
- Normal model, $\lambda_j \in \{1, 2, \ldots, 5\}$, need $\gamma_1 > 1/(2\lambda_{\text{min}}) = 0.5$. 
Optimal SGD: transformation

- Ideally, $\mathcal{I}(\theta) = I$. In statistics $\theta$ is called *location parameter*.
- SGD works great with location params. – could transform to get there.

Example: Suppose that $Y \sim N(\theta, 1)$ and we wish to infer $\theta$; let $\gamma_n = \frac{1}{n}$.

Vanilla SGD iterations:

$$
\theta_n = \theta_{n-1} + \frac{3}{n} \theta_n - \theta_{n-1} \left( y_n - \theta_{n-1} \right).
$$

Transformation:

Set $\mu = \theta_3$, then run SGD on transformed space:

$$
\mu_n = \mu_{n-1} + \frac{1}{n} \left( y_n - \mu_{n-1} \right), \quad \theta_n = \mu_n^{1/3}.
$$

Here, $I(\theta) = 15 \theta^4 \equiv \psi > 0$; but, $I(\mu) = I(\theta) (\frac{d\theta}{d\mu})^2 = 1$.

For vanilla SGD,

$$
n \text{Var}(\theta_n) \rightarrow \frac{\psi}{2\psi - 1}.
$$

For transformed,

$$
n \text{Var}(\mu_n) \rightarrow 1 \quad \text{and} \quad n \text{Var}(\theta_n) \rightarrow \frac{1}{\psi} = I(\theta) - 1 – \text{optimal!}
$$

Unfortunately, general case is hard: find transformation with Jacobian $J$ s.t. $J I(\theta) J^\top \approx \text{const.}$
Optimal SGD: transformation

- Ideally, $\mathcal{I}(\theta) = \mathbb{I}$. In statistics $\theta$ is called location parameter.
- SGD works great with location params. – could transform to get there.

Example:

Suppose that $Y \sim \mathcal{N}(\theta^3, 1)$ and we wish to infer $\theta$; let $\gamma_n = 1/n$.

Vanilla SGD iterations:

$$
\theta_n = \theta_{n-1} + \frac{3}{n} \theta_{n-1}^2 (y_n - \theta_{n-1}^3).
$$

**Transformation:** Set $\mu = \theta^3$, then run SGD on transformed space:

$$
\mu_n = \mu_{n-1} + \frac{1}{n} (y_n - \mu_{n-1}), \quad \theta_n = \mu_n^{1/3}. \text{(i.e., } \mu_n = \bar{y}_n)\]
Optimal SGD: transformation

- Ideally, $\mathcal{I}(\theta) = \mathbb{I}$. In statistics $\theta$ is called \textit{location parameter}.
- SGD works great with location params. – could transform to get there.

\textbf{Example:}

Suppose that $Y \sim \mathcal{N}(\theta^3, 1)$ and we wish to infer $\theta$; let $\gamma_n = 1/n$. Vanilla SGD iterations:

$$\theta_n = \theta_{n-1} + \frac{3}{n} \theta_{n-1}^2 (y_n - \theta_{n-1}^3).$$

\textbf{Transformation:} Set $\mu = \theta^3$, then run SGD on transformed space:

$$\mu_n = \mu_{n-1} + \frac{1}{n} (y_n - \mu_{n-1}), \quad \theta_n = \mu_{n}^{1/3} \text{ (i.e., } \mu_n = \bar{y}_n)$$

- Here, $\mathcal{I}(\theta) = 15\theta^4 \equiv \psi > 0$; but, $\mathcal{I}(\mu) = \mathcal{I}(\theta) \left( \frac{d\theta}{d\mu} \right)^2 = 1$. 
Ideally, \( I(\theta) = \mathbb{I} \). In statistics \( \theta \) is called \textit{location parameter}. SGD works great with location params. – could transform to get there.

Example:

Suppose that \( Y \sim \mathcal{N}(\theta^3, 1) \) and we wish to infer \( \theta \); let \( \gamma_n = 1/n \).

Vanilla SGD iterations:

\[
\theta_n = \theta_{n-1} + \frac{3}{n} \theta_{n-1}^2 (y_n - \theta_{n-1}^3).
\]

Transformation: Set \( \mu = \theta^3 \), then run SGD on transformed space:

\[
\mu_n = \mu_{n-1} + \frac{1}{n} (y_n - \mu_{n-1}), \quad \theta_n = \mu_{n}^{1/3}. \text{(i.e., } \mu_n = \bar{y}_n \text{)}
\]

Here, \( I(\theta) = 15\theta^4 \equiv \psi > 0 \); but, \( I(\mu) = I(\theta)(\frac{d\theta}{d\mu})^2 = 1 \).

For vanilla SGD, \( n\text{Var}(\theta_n) \rightarrow \psi/(2\psi - 1) \).
Optimal SGD: transformation

- Ideally, \( \mathcal{I}(\theta) = \mathbb{I} \). In statistics \( \theta \) is called \textit{location parameter}.
- SGD works great with location params. – could transform to get there.

**Example:**

Suppose that \( Y \sim \mathcal{N}(\theta^3, 1) \) and we wish to infer \( \theta \); let \( \gamma_n = 1/n \).

Vanilla SGD iterations:

\[
\theta_n = \theta_{n-1} + \frac{3}{n} \theta_{n-1}^2 (y_n - \theta_{n-1}^3).
\]

**Transformation:** Set \( \mu = \theta^3 \), then run SGD on transformed space:

\[
\mu_n = \mu_{n-1} + \frac{1}{n} (y_n - \mu_{n-1}), \quad \theta_n = \mu_n^{1/3}. \text{(i.e., } \mu_n = \bar{y}_n) \]

- Here, \( \mathcal{I}(\theta) = 15\theta^4 \equiv \psi > 0 \); but, \( \mathcal{I}(\mu) = \mathcal{I}(\theta) \left( \frac{d\theta}{d\mu} \right)^2 = 1 \).
- For vanilla SGD, \( n \text{Var}(\theta_n) \to \psi/(2\psi - 1) \).
- For transformed, \( n \text{Var}(\mu_n) \to 1 \) and \( n \text{Var}(\theta_n) \to 1/\psi = \mathcal{I}(\theta)^{-1} \) – optimal!
Ideally, $\mathcal{I}(\theta) = \mathbb{I}$. In statistics $\theta$ is called \textit{location parameter}.

SGD works great with location params. – could transform to get there.

**Example:**

Suppose that $Y \sim \mathcal{N}(\theta^3, 1)$ and we wish to infer $\theta$; let $\gamma_n = 1/n$.

Vanilla SGD iterations:

$$\theta_n = \theta_{n-1} + \frac{3}{n} \theta^2_{n-1} (y_n - \theta^3_{n-1}).$$

**Transformation:** Set $\mu = \theta^3$, then run SGD on transformed space:

$$\mu_n = \mu_{n-1} + \frac{1}{n} (y_n - \mu_{n-1}), \quad \theta_n = \mu_n^{1/3}. \text{ (i.e., } \mu_n = \bar{y}_n)$$

Here, $\mathcal{I}(\theta) = 15\theta^4 \equiv \psi > 0$; but, $\mathcal{I}(\mu) = \mathcal{I}(\theta) \left( \frac{d\theta}{d\mu} \right)^2 = 1$.

For vanilla SGD, $n \text{Var}(\theta_n) \rightarrow \psi/(2\psi - 1)$.

For transformed, $n \text{Var}(\mu_n) \rightarrow 1$ and $n \text{Var}(\theta_n) \rightarrow 1/\psi = \mathcal{I}(\theta)^{-1}$ – optimal!

Unfortunately, general case is hard: find transformation with Jacobian $J$ s.t.

$$J\mathcal{I}(\theta) J^\top \approx \text{const.}$$
Simulation

- Suppose that $Y \sim \mathcal{N}(\theta^3, 20^2)$ and $\theta = 4.5$. 

![Graph showing simulation results with red line labeled 'transformed' and black line labeled 'vanilla (implicit)'](image-url)
Asymptotic normality

- Under typical Lindeberg conditions,

\[ \sqrt{n}(\theta_n^\text{im} - \theta^*) \to \mathcal{N}(0, \Sigma_{\theta^*}, \gamma_1). \]
Asymptotic normality

- Under typical Lindeberg conditions,

\[ \sqrt{n}(\theta^{im}_n - \theta^*_n) \to \mathcal{N}(0, \Sigma_{\theta^*}, \gamma_1). \]

Figure: Experiment with \( y \sim \mathcal{N}(\theta^3, \sigma^2) \). **Left:** \( n \text{Var}(\theta_n) \), for \( n = 1, \ldots, N = 3e4 \), where var is over 250 samples (green = \( \gamma_1/2 \)); **right:** std. \( (\theta_N - \theta^*_n)/(\sqrt{\gamma_1/2N}) \).
Asymptotic normality: multivariate example

Figure: Samples of $(\theta_n - \theta_\star)^T \Sigma_{\theta_\star, \gamma_1}^{-1} (\theta_n - \theta_\star)$. 
1. Unconditional stability.
2. Quantifiable efficiency loss (+optimality).
3. Asymptotic normality.
Summing up the good properties of implicit SGD

1. Unconditional stability.
2. Quantifiable efficiency loss (+optimality).
3. Asymptotic normality.

Two main talk points:
- Asymptotic variance of SGD in closed form; statistical implications.
- Asymptotic normality, especially of implicit SGD thanks to its stability.
Ongoing work

- Statistics of optimization procedures (e.g., Second-order procedures).
- Implicit SGD + constant rates to speed-up convergence (w/ Jerry Chee).
- Networks/intractable likelihoods (e.g., Monte-Carlo SGD, w/ Mladen).
- **Future**: Reinforcement learning (esp. dynamic choice models); neural networks.
References

- D Tran, PT, Edoardo M. Airoldi, ”sgd R package: stochastic gradient methods for estimation with large data sets” (2016, Journal of Statistical Software; minor rev.)
- PT, D Tran, EM Airoldi, ”Towards stability and optimality in stochastic gradient descent” (2016, AI & Statistics, AISTATS)
\[ \theta_* = (0.37, 0.15)^\top, \quad \lambda_{\text{min}} = 1, \quad \lambda_{\text{max}} = 1, \quad \gamma_1 \in \{0.1, 1, 10\}. \]
\( \theta_\star = (0.37, 0.15)^T, \lambda_{\text{min}} = 1, \lambda_{\text{max}} = 10, \gamma_1 \in \{0.1, 1, 10\}. \)
Intuition: implicit update as an infinite series of standard updates:

\[ \theta(0)_n = \theta(n) - \frac{1}{n} \left( Y_n - e^{\theta(n-1)} \right) \]
\[ \theta(1)_n = \theta(n) - \frac{1}{n} \left( Y_n - e^{\theta(0)_n} \right) \]
\[ \theta(2)_n = \theta(n) - \frac{1}{n} \left( Y_n - e^{\theta(1)_n} \right) \]

\[ \cdots \]
\[ \theta(\infty)_n = \theta(n) - \frac{1}{n} \left( Y_n - e^{\theta(\infty)} \right) \]

cf. self-consistency \( \Delta \) principle in statistics (Efron, 1967); (Tarpey & Flury, 1996).
Intuition: implicit update as an \textbf{infinite} series of standard updates:

\[
\theta_n^{(0)} = \theta_{n-1} + \frac{1}{n} (Y_n - e^{\theta_{n-1}}).
\]
Intuition: implicit update as an infinite series of standard updates:

\[ \theta^{(0)}_n = \theta_{n-1} + \frac{1}{n} (Y_n - e^{\theta_{n-1}}). \]

\[ \theta^{(1)}_n = \theta_{n-1} + \frac{1}{n} (Y_n - e^{\theta^{(0)}_n}). \]
Intuition: implicit update as an infinite series of standard updates:

\[
\theta_{n}^{(0)} = \theta_{n-1} + \frac{1}{n}(Y_n - e^{\theta_{n-1}}).
\]

\[
\theta_{n}^{(1)} = \theta_{n-1} + \frac{1}{n}(Y_n - e^{\theta_{n}^{(0)}}).
\]

\[
\theta_{n}^{(2)} = \theta_{n-1} + \frac{1}{n}(Y_n - e^{\theta_{n}^{(1)}}).
\]

Intuition: implicit update as an \textbf{infinite} series of standard updates:

\[ \theta^{(0)}_n = \theta_{n-1} + \frac{1}{n} (Y_n - e^{\theta_{n-1}}). \]

\[ \theta^{(1)}_n = \theta_{n-1} + \frac{1}{n} (Y_n - e^{\theta^{(0)}_n}). \]

\[ \theta^{(2)}_n = \theta_{n-1} + \frac{1}{n} (Y_n - e^{\theta^{(1)}_n}). \]

\[ \cdots \]

\[ \theta^{(\infty)}_n = \theta_{n-1} + \frac{1}{n} (Y_n - e^{\theta^{(\infty)}_n}). \]

Efficient computation of implicit updates

Suppose $Y_n \sim \text{Poisson}(e^{X_n^T \theta_*})$. Then,
Efficient computation of implicit updates

Suppose $Y_n \sim \text{Poisson}(e^{X_n^T \theta_*})$. Then,

$$
\theta_{im}^n = \theta_{im}^{n-1} + \gamma_n(Y_n - e^{X_n^T \theta_{im}^n})X_n
$$

(1)

$$
= \theta_{im}^{n-1} + \gamma_n \xi_n(Y_n - e^{X_n^T \theta_{im}^{n-1}})X_n
$$

(2)

$$
\triangleq \theta_{im}^{n-1} + a_n X_n.
$$

(3)
Efficient computation of implicit updates

Suppose $Y_n \sim \text{Poisson}(e^{X_n^T \theta^*})$. Then,

$$
\theta_{n}^{\text{im}} = \theta_{n-1}^{\text{im}} + \gamma_n (Y_n - e^{X_n^T \theta_{n}^{\text{im}}}) X_n \tag{1}
$$

$$
= \theta_{n-1}^{\text{im}} + \gamma_n \xi_n (Y_n - e^{X_n^T \theta_{n-1}^{\text{im}}}) X_n \tag{2}
$$

$$
\triangleq \theta_{n-1}^{\text{im}} + a_n X_n. \tag{3}
$$

Equate the two scales:

$$
a_n = \gamma_n (Y_n - e^{X_n^T \theta_{n}^{\text{im}}}) \quad \text{[by setting (1) = (3)]}
$$

$$
= \gamma_n (Y_n - e^{X_n^T \theta_{n-1}^{\text{im}} + \|X_n\|^2 a_n}). \quad \text{[by substituting $\theta_{n}^{\text{im}}$ with (3)]}
$$

LHS $\uparrow a_n$ and RHS $\downarrow a_n$, both convex. Fixed-point equation is

$$
x = a - be^{cx},
$$

where $b, c > 0$. It follows that $x \in [\min(0, a - b), \max(0, a - b)]$. 

Back to main.
Self-consistency principle

Example. Estimate CDF $F(t)$ with data $Y_1, Y_2, \ldots, Y_n$; $Y^{obs} = \text{uncensored}$. 
**Example.** Estimate CDF $F(t)$ with data $Y_1, Y_2, \ldots, Y_n$; $Y_{obs} =$ uncensored.

A self-consistent estimator of $F(t)$ is

$$F^*(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E} \left( \mathbb{I}\{Y_i \leq t\} | Y_{obs}, F^* \right).$$
Stochastic approximation

- In an experiment, suppose $\theta$ is input, $H(\theta)$ random output.
- Suppose we wish to find $\theta^*$ such that

$$E(H(\theta^*)) = 0.$$
In an experiment, suppose $\theta$ is input, $H(\theta)$ random output.
Suppose we wish to find $\theta_\star$ such that
\[
\mathbb{E}(H(\theta_\star)) = 0.
\]
Robbins-Monro (1951) stochastic approximation procedure:
\[
\theta_n = \theta_{n-1} + \gamma_n H(\theta_{n-1}).
\]
Theorem (Robbins and Monro, 1951): $\mathbb{E}(\big|\theta_n - \theta_\star\big|^2) \to 0$ if
\begin{itemize}
\item $\sum \gamma_i = \infty$; $\sum \gamma_i^2 < \infty$;
\item $H$ is concave in expectation and Lipschitz;
\item $\mathbb{E}(\big|\big|H(\theta_\star)\big|\big|^2) < \infty$.
\end{itemize}
SGD as special case: $H(\theta) \equiv \nabla \log f(Y; X, \theta)$ and $\theta_n \to \theta_\star$ because
\[
\mathbb{E}(\nabla \log f(Y; X, \theta_\star)) = 0.
\]
Implicit stochastic approximation

- Classical stochastic approximation of Robbins & Monro (1951)
  \[ \theta_n = \theta_{n-1} + \gamma_n H(\theta_{n-1}) \]

- **Implicit** stochastic approximation (Toulis & Airoldi, 2015b)
  \[ \theta_n = \theta_{n-1} + \gamma_n H(\theta^*_{n-1}) \]
  \[ \text{s.t. } \mathbb{E}(\theta_n|\theta_{n-1}) = \theta^*_{n-1} \]

- Non-asymptotic/asymptotic analysis (Toulis & Airoldi, 2015b)
- Implementations need to estimate \( \theta^*_{n-1} \)
Consider the second-order implicit SGD procedure

$$\theta_{n}^{\text{im}} = \theta_{n-1}^{\text{im}} + \frac{1}{n} C_n \nabla \log f(Y_n; X_n, \theta_{n}^{\text{im}}),$$

where $C_n \to C > 0$, where $C$ is symmetric and commutes with $\mathcal{I}(\theta^\star)$. Then

$$n \text{Var}(\theta_{n}^{\text{im}}) \to (2C\mathcal{I}(\theta^\star) - I)^{-1} C\mathcal{I}(\theta^\star) C \triangleq \Sigma_{\theta^\star,C}.$$
Theorem (Toulis & Airoldi, 2015a)

Consider the second-order implicit SGD procedure

\[ \theta_{n}^{\text{im}} = \theta_{n-1}^{\text{im}} + \frac{1}{n} C_n \nabla \log f(Y_n; X_n, \theta_{n}^{\text{im}}), \]

where \( C_n \to C > 0, \) where \( C \) is symmetric and commutes with \( \mathcal{I}(\theta_\star) \). Then

\[ n \text{Var}(\theta_{n}^{\text{im}}) \to (2C \mathcal{I}(\theta_\star) - \mathbb{I})^{-1} C \mathcal{I}(\theta_\star) C \triangleq \Sigma_{\theta_\star, C}. \]

- Optimal efficiency **only** if \( C = \mathcal{I}(\theta_\star)^{-1} \).
Optimal efficiency: second-order SGD

Theorem (Toulis & Airoldi, 2015a)

Consider the second-order implicit SGD procedure

\[ \theta_{n}^{\text{im}} = \theta_{n-1}^{\text{im}} + \frac{1}{n} C_{n} \nabla \log f(Y_{n}; X_{n}, \theta_{n}^{\text{im}}), \]

where \( C_{n} \to C > 0 \), where \( C \) is symmetric and commutes with \( \mathcal{I}(\theta_{\star}) \).

Then

\[ n \text{Var}(\theta_{n}^{\text{im}}) \to (2C\mathcal{I}(\theta_{\star}) - \mathbb{I})^{-1} C\mathcal{I}(\theta_{\star})C \triangleq \Sigma_{\theta_{\star},C}. \]

- Optimal efficiency **only** if \( C = \mathcal{I}(\theta_{\star})^{-1} \).
- **Adaptive** methods concurrently estimate \( \mathcal{I}(\theta_{\star})^{-1} \);
  e.g., \( C_{n} = \mathcal{I}(\theta_{n-1})^{-1} \), Sakrison’s (1965) explicit procedure.

Back to [main](#). Compare with [AdaGrad](#). See also implicit method with [averaging](#).
A note on AdaGrad

A popular adaptive procedure is AdaGrad (Duchi et.al., 2011)

$$\theta_{n}^{\text{ada}} = \theta_{n-1}^{\text{ada}} + \gamma_1 \frac{1}{\sqrt{n}} C_n^{1/2} \nabla \log f(Y_n; X_n, \theta_{n-1}^{\text{ada}}),$$

where $C_n \rightarrow \text{diag}(I(\theta^*)^{-1}).$
A note on AdaGrad

- A popular adaptive procedure is AdaGrad (Duchi et.al., 2011)

\[
\theta_n^{\text{ada}} = \theta_{n-1}^{\text{ada}} + \gamma_1 \frac{1}{\sqrt{n}} C_n^{1/2} \nabla \log f(Y_n; X_n, \theta_{n-1}^{\text{ada}}),
\]

where \( C_n \to \text{diag}(\mathcal{I}(\theta^*))^{-1} \).

(Toulis & Airoldi, 2015a)

\[
\sqrt{n} \text{Var}(\theta_n^{\text{ada}}) \to \frac{\gamma_1}{2} \text{diag}(\mathcal{I}(\theta^*))^{-1/2}. \tag{4}
\]

- AdaGrad is inefficient but (1) holds regardless of \( \gamma_1 \).
- In contrast, SGD procedures require \( \gamma_1 > 1/(2\lambda_{\min}) \) for \( O(1/n) \) efficiency.
\[ \theta_* = (2.23, 0.5, 0.1, 0.02, 0.01)^T; \quad \lambda_j \in [1, 10] \]
Explicit stochastic approximation: implementations

\[ \theta_n = \theta_{n-1} + \gamma_n H(\theta_{n-1}^*) \]

s.t. \[ \mathbb{E}(\theta_n | \theta_{n-1}) = \theta_{n-1}^* \]

1. Run separate RM procedure at each \( n \)th iteration, \( k = 1, 2, \ldots \)

\[ x_k = x_{k-1} + a_k [\theta_{n-1} + \gamma_n H(x_{k-1}) - x_{k-1}] \]

- \( x_k \to \theta_{n-1}^* \) (few iterations of \( x_k \) can be enough)
- Only choice if can only sample through \( H \) (classical RM)
- Related to “multiple timescales” (Borkar, 2009)

2. Use \( \theta_n \) as an estimate of \( \theta_{n-1}^* \) ! Results in familiar procedure

\[ \theta_n = \theta_{n-1} + \gamma_n H(\theta_n) \]

- Possible if \( H \) is known in analytic form (as in implicit SGD)
Asymptotic optimal efficiency: averaging

Theorem (Toulis et.al., 2016)

Consider the averaged procedure, where $\gamma_n \propto n^{-\gamma}$, $\gamma \in (0, 1)$, $\lambda_{\min} > 0$,

$$
\theta_{n}^{im} = \theta_{n-1}^{im} + \gamma_n \nabla \log f(Y_n; X_n, \theta_{n}^{im})
$$

$$
\bar{\theta}_n = \frac{1}{n} \sum_{i=1}^{n} \theta_{i}^{im}.
$$

Then, $\bar{\theta}_n$ has asymptotically optimal efficiency, i.e.,

$$
n \text{Var}(\bar{\theta}_n) \rightarrow \mathcal{I}(\theta^\star)^{-1}.
$$

- $\lambda_{\min} > 0$ critical for theorem; typically, $\gamma_n \propto 1/\sqrt{n}$.
- Classical averaging results: (Ruppert, 1988); (Bather, 1989); (Polyak & Juditsky, 1992)

Back to Second-order efficiency result. 

Panagiotis (Panos) Toulis

Statistical properties of stochastic gradient methods
Implicit SGD can be written as

$$\theta_{im}^n = \arg \max_{\theta} \left\{ \log f(Y_n; X_n, \theta) - \frac{1}{2\gamma_n} ||\theta - \theta_{im}^{n-1}||^2 \right\}.$$ 

Thus, $\theta_{im}^n$ is the \textit{posterior mode} of the Bayesian model,

$$\theta | \theta_{im}^{n-1} \sim \mathcal{N}(\theta_{im}^{n-1}, \gamma_n I)$$

$$Y_n | X_n, \theta \sim f$$

- Implicit SGD: interpretation of $\gamma_n$ as information parameter.
- Explicit SGD: interpretation of $\gamma_n$ as “step-size”.

First implicit method by Nagumo & Noda (1967); (Slock, 1993)
Connection to proximal methods

- In optimization problem, $\text{arg min}_\theta g(\theta)$, for deterministic $g$ we can do

$$\theta_n = \text{arg min}_\theta \left\{ g(\theta) + \frac{1}{2\gamma_n} ||\theta - \theta_{n-1}||^2 \right\}.$$  

- RHS is a proximal operator, say $\text{prox}_{\gamma_n g}(\theta_{n-1})$.

- Stochastic proximal procedures (Duchi et.al., 2009); (Rosasco et.al., 2014):

$$\theta_n = \text{prox}_{\gamma_n R} (\theta_{n-1} + \gamma_n \nabla \log f(Y_n; X_n, \theta_{n-1}))$$

- $R$ is a deterministic regularizer; in implicit SGD it is random.

- Such methods make one explicit step and then one deterministic proximal step (implicit update). May be unstable.

Back to related work.
Consider the problem

\[ \hat{\theta} = \arg \min_{\theta} \sum_{i=1}^{N} f_i(\theta). \]

where \( N = \# \) datapoints, \( i = \) datapoint index, \( f_i = \) loss at \( i \) datapoint.

Bertsekas (2011) analyzed the procedure

\[ \theta_n = \arg \min_{\theta} \left\{ f_{i_n}(\theta) + \frac{1}{2\gamma_n} \| \theta - \theta_{n-1} \|^2 \right\}, \]

where \( i_n \in \{1, 2, \ldots, N\} \).

Like implicit SGD but in a non-streaming setting (fixed dataset).

Analysis compares \( i_n \) cycling through data with random \( i_n \).

Back to related work.
Optimal rates: a surprising pivotal quantity

- One *principled* way to set the optimal rate:

\[
\gamma_1^* = \arg\min_{\gamma_1} \text{tr}(\Sigma_{\theta_*}, \gamma_1) \iff \gamma_1^* = \arg\min_{\gamma_1} \sum_{j=1}^{p} \frac{\gamma_1^2 \lambda_j}{2\gamma_1 \lambda_j - 1}.
\]

- If \( \gamma_1 >> 1/(2\lambda_{\text{min}}) \),

\[
\text{tr}(\Sigma_{\theta_*}, \gamma_1) \approx p \frac{\gamma_1}{2}. \text{ In fact, } \Sigma_{\theta_*}, \gamma_1 \approx \frac{\gamma_1}{2} \mathbb{I} \text{ (parameter-free!)}
\]

- Fairly general way to construct pivotal quantity for \( \theta_* \).

- But we pay price in efficiency.

Back to [optimal rates](#).
The unusual technical challenge of implicit SGD

- Standard asymptotic analysis obtains recursion for $\mathbb{E} \left( \| \theta_n^{\text{ex}} - \theta^* \|^2 \right)$. 

However, in the implicit procedure $\theta_n^{\text{im}} = \theta_n^{\text{im}} - 1 + \gamma_n \nabla \log f(Y_n; X_n, \theta_n^{\text{im}})$, we cannot use standard analysis because $(Y_n, X_n) \not\perp \perp \theta_n^{\text{im}}$. 

The unusual technical challenge of implicit SGD

- Standard asymptotic analysis obtains recursion for $\mathbb{E} \left( \| \theta_n^{ex} - \theta^* \|^2 \right)$.
- A crucial property is the concavity of

$$\mathbb{E} \left( \nabla \log f(Y_n; X_n, \theta_{n-1}^{ex}) | \theta_{n-1}^{ex} \right),$$

which requires

$$(Y_n, X_n) \perp \perp \theta_{n-1}^{ex}.$$
The unusual technical challenge of implicit SGD

- Standard asymptotic analysis obtains recursion for $\mathbb{E} \left( \| \theta_n^{\text{ex}} - \theta_* \|^2 \right)$.
- A crucial property is the concavity of $\mathbb{E} \left( \nabla \log f(Y_n; X_n, \theta_n^{\text{ex}}) | \theta_n^{\text{ex}} \right)$, which requires $(Y_n, X_n) \perp \perp \theta_n^{\text{ex}}$.

- However, in the implicit procedure
  
  $$\theta_n^{\text{im}} = \theta_{n-1}^{\text{im}} + \gamma_n \nabla \log f(Y_n; X_n, \theta_n^{\text{im}})$$

  we cannot use standard analysis because $(Y_n, X_n) \not\perp \theta_n^{\text{im}}$. 

Unusual technical challenge: our approach

- In many statistical models

\[ f(Y; X, \theta) \equiv f(Y; X, X^\top \theta). \]
Unusual technical challenge: our approach

- In many statistical models

\[ f(Y; X, \theta) \equiv f(Y; X, X^\top \theta). \]

- Then, \( \nabla \log f(Y; X, \theta) \) collinear with \( X \) (free of \( \theta \)); thus,

\[
\theta_n^{\text{im}} = \theta_{n-1}^{\text{im}} + \gamma_n \nabla \log f(Y_n; X_n, \theta_n^{\text{im}}) \\
= \theta_{n-1}^{\text{im}} + \gamma_n \xi_n \nabla \log f(Y_n; X_n, \theta_{n-1}^{\text{im}}). 
\]
Unusual technical challenge: our approach

- In many statistical models

\[ f(Y; X, \theta) \equiv f(Y; X, X^\top \theta). \]

- Then, \( \nabla \log f(Y; X, \theta) \) collinear with \( X \) (free of \( \theta \)); thus,

\[
\theta_{n}^{im} = \theta_{n-1}^{im} + \gamma_n \nabla \log f(Y_n; X_n, \theta_{n}^{im}) \\
= \theta_{n-1}^{im} + \gamma_n \xi_n \nabla \log f(Y_n; X_n, \theta_{n-1}^{im}).
\]

1. \( \xi_n \) is easy to calculate \( \Rightarrow \) fast implementation!
2. a.s. bound for \( \xi_n \) \( \Rightarrow \) avoids conditioning problem since

\( (Y_n, X_n) \perp \perp \theta_{n-1}^{im}. \)

Proceed with analysis. Back to main.
Almost-sure bound for $\xi_n$

- Start with

$$\theta^\text{im}_n = \theta^\text{im}_{n-1} + \gamma_n \xi_n \nabla \log f(Y_n; X_n, \theta^\text{im}_{n-1}).$$
Almost-sure bound for $\xi_n$

- Start with
  \[
  \theta_{n}^{\text{im}} = \theta_{n-1}^{\text{im}} + \gamma_n \xi_n \nabla \log f(Y_n; X_n, \theta_{n-1}^{\text{im}}).
  \]

- Let $\hat{I}(\theta) = -\nabla^2 \log f(Y; X, \theta)$ and suppose $\text{tr}(\hat{I}(\theta)) \geq s > 0$.

- Then, Taylor expansion of gradient around $\theta_{n-1}^{\text{im}}$ yields
  \[
  \xi_n \geq (1 + \gamma_n s)^{-1} \text{ a.s.}
  \]
Almost-sure bound for $\xi_n$

- Start with

$$\theta_n^{im} = \theta_{n-1}^{im} + \gamma_n \xi_n \nabla \log f(Y_n; X_n, \theta_{n-1}^{im}).$$

- Let $\hat{I}(\theta) = -\nabla^2 \log f(Y; X, \theta)$ and suppose $\text{tr}(\hat{I}(\theta)) \geq s > 0$.

- Then, Taylor expansion of gradient around $\theta_{n-1}^{im}$ yields

$$\xi_n \geq (1 + \gamma_n s)^{-1} \text{ a.s.}$$

- Now, $(X_n, Y_n) \perp \perp \theta_{n-1}^{im}$ yields recursion for MSE,

$$\mathbb{E} \left( \left\| \theta_n^{im} - \theta^* \right\|^2 \right) \leq \frac{1}{1 + \gamma_n s} \mathbb{E} \left( \left\| \theta_{n-1}^{im} - \theta^* \right\|^2 \right) + O(\gamma_n^2).$$

Back to main. Proceed to solving the recursion.
The wonderful idea of majorization-minorization

Suppose we wish to solve $b_n \leq F(b_{n-1})$, $F$ non-decreasing.

(\textbf{majorize}) Instead, we solve $c_n^\alpha \geq F(c_{n-1}^\alpha)$. If $b_0 \leq c_0^\alpha$ then

$$b_1 \leq F(b_0) \leq F(c_0^\alpha) \leq c_1^\alpha \Rightarrow b_n \leq c_n^\alpha. \text{ (by induction)}$$

(\textbf{minorize}) Minimize $c_n^*$ wrt $\alpha$ to min. upper bound, $b_n \leq c_n^*$. 
Suppose we wish to solve $b_n \leq b_{n-1} + n$, $b_0 = 0$. Clearly, the solution is

$$b_n \leq 1 + 2 + \ldots + n \leq n(n + 1)/2.$$ 

But suppose we don’t know the correct form but suspect it is $\alpha_0 n^2 + \alpha_1 n$. 

Thus, $\alpha_0 \geq 5$ and $\alpha_1 \geq \alpha_0$. Therefore, $b_n \leq c^* n = \arg \min \alpha c_\alpha n = n(n + 1)/2$. 

Back to main ⊿

Panagiotis (Panos) Toulis

Statistical properties of stochastic gradient methods 45 / 24
The wonderful idea of majorization-minorization

A simple example

Suppose we wish to solve \( b_n \leq b_{n-1} + n, \ b_0 = 0 \). Clearly, the solution is

\[
b_n \leq 1 + 2 + \ldots + n \leq n(n+1)/2.
\]

But suppose we don’t know the correct form but suspect it is \( \alpha_0 n^2 + \alpha_1 n \). Then define \( c_n^\alpha = \alpha_0 n^2 + \alpha_1 n \) and solve:

\[
c_n^\alpha \geq c_{n-1}^\alpha + n
\]

\[
\alpha_0 n^2 + \alpha_1 n \geq \alpha_0 (n-1)^2 + \alpha_1 (n-1) + n
\]

\[
(2\alpha_0 - 1)n + \alpha_1 \geq \alpha_0
\]

Thus, \( \alpha_0 \geq .5 \) and \( \alpha_1 \geq \alpha_0 \). Therefore,

\[
b_n \leq c_n^* = \arg \min_{\alpha} c_n^\alpha = .5n^2 + .5n = n(n+1)/2
\]
In many cases the likelihood is intractable, thus SGD cannot be used.
In many cases the likelihood is intractable, thus SGD cannot be used.

Suppose finite data, and take $S^{obs}$ to be the sufficient statistic.

Define $T(\theta) = \mathbb{E} (S|\theta)$, e.g., through Monte-Carlo.
In many cases the likelihood is intractable, thus SGD cannot be used.

Suppose finite data, and take $S^{obs}$ to be the sufficient statistic.

Define $T(\theta) = \mathbb{E}(S|\theta)$, e.g., through Monte-Carlo.

Then calculate the update,

$$\theta_n = \theta_{n-1} + \gamma_n (S^{obs} - T(\theta_{n-1})),$$

For instance, $S^{obs}$ observed network statistics (e.g., #triangles), $T =$ simulated average statistics.

By SA theory $\theta_n$ converges to point $\theta_\infty$ such that

$$T(\theta_\infty) = S^{obs}.$$