Review of stochastic optimization with emphasis on common models and tasks in econometrics

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

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

Modern (and some classical) estimation problems are computationally hard. Need for stochastic (approximate) methods of optimization.
Optimization and estimation are complementary.
Modern (and some classical) estimation problems are computationally hard. Need for stochastic (approximate) methods of optimization.
A unique method that crosses the optimization-statistics boundary:

**stochastic gradient descent (SGD).**

Special case of *stochastic approximation* (Robbins & Monro, 1951).
Robbins-Monro procedure (1951)

- Suppose we wish to solve the problem:

\[
\text{find } \theta_\star \text{ s.t. } \mathbb{E}(H(\theta_\star)) = 0.
\]

- \(H(\theta)\) is random output given \(\theta\) (may have unknown form).
- Robbins-Monro proposed the following estimation procedure:

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

- Conditions (roughly):
  - \(\sum \gamma_i = \infty \) and \(\sum \gamma_i^2 < \infty\) (for example, \(\gamma_n \propto 1/n\));
  - \(\mathbb{E}(H(\theta)) \equiv h(\theta) = \text{convex}\).
  - \(||H(\theta)||^2\) is bounded almost-surely.

- Under such conditions,

\[
\theta_n \rightarrow \theta_\star, \text{ in probability.}
\]
Stochastic gradient descent (SGD)

- 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.
Stochastic gradient descent (SGD)

- 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.

- 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).
Stochastic gradient descent (SGD)

- 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:

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

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

- To connect to RM procedure, set \( H(\theta) = \nabla l_J(\theta) \) with

\[
\mathbb{E}(H(\theta)) = \frac{1}{N} \sum_{i=1}^{N} l_i(\theta).
\]
Streaming-data version of SGD

- For simplicity, consider a stream of iid $((X_i, Y_i) \sim f_{\theta^*}, i = 1, 2, \ldots$.
- The SGD estimator of $\theta^*$ after $n$ datapoints is

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

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

Stochastic approximation theory:

$$\mathbb{E} (\nabla \log f(Y; X, \theta_\infty)) = 0 \iff \theta_\infty = \theta^*.$$
Streaming-data version of SGD

- For simplicity, consider a stream of iid \((X_i, Y_i) \sim f_{\theta\star}, i = 1, 2, \ldots\)
- The **SGD estimator** of \(\theta\star\) after \(n\) datapoints is

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

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

**Stochastic approximation theory:**

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

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

possible datasets of size $n$

sampling distribution:
$E|\theta_n - \theta_*|, \text{Var}(\theta_n)$
Example on linear 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 matrix:

$$I(\theta) = -\mathbb{E}(\nabla^2 \log f(Y; X, \theta)) = \mathbb{E}(XX^\top)$$

Crucial for performance of SGD.
Example on linear normal model

- Suppose $Y | X \sim \mathcal{N}(X^T \theta_\star, 1)$. Then, $\nabla \log f(Y; X, \theta) = (Y - X^T \theta) X$.

- SGD estimator:

  $$
  \theta_n = \theta_{n-1} + \gamma_n (Y_n - X_n^T \theta_{n-1}) X_n
  = (I - \gamma_n X_n X_n^T) \theta_{n-1} + \gamma_n Y_n X_n.
  \tag{1}
  $$

- Fisher information matrix: $\mathcal{I}(\theta) = -\mathbb{E} (\nabla^2 \log f(Y; X, \theta)) = \mathbb{E} (XX^T)$.

  - Crucial for performance of SGD.

- Let $\lambda_{\min} = \min \text{eig } \mathbb{E} (XX^T)$, and let $\lambda_{\max}$ be the max. eigenvalue.
\( \theta_\star = (0.47, 0.22, 0.1, \ldots)^\top \in \mathbb{R}^{20}, \lambda_{\min} = \lambda_{\max} = 1, \gamma_1 = 1. \)
\( \theta_* = (0.47, 0.22, 0.1, \ldots)^T \in \mathbb{R}^{20}, \lambda_{\text{min}} = \lambda_{\text{max}} = 1, \gamma_1 = 10. \)
In classical SGD, solve recursively for $\theta_n$ to derive:

$$\theta_n = P^n_1 \theta_0 + \sum_{i=1}^{n} \gamma_i P^n_{i+1} Y_i X_i,$$

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

Discounts initial conditions $\theta_0$. Its "size" is critical for stability:

$$\prod_{j=1}^{n} (1 - \gamma_1 \lambda_j)^{1 - \gamma_1 2 \lambda_j},$$

where $\lambda_j$ is the $j$th eigenvalue of $E(XX^T)$.

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

$$\max_{n>0} \{\text{eig}(E(P^n_1))\} = O\left(2^{\gamma_1 \lambda_{\text{max}}} / \sqrt{\gamma_1 \lambda_{\text{max}}}ight).$$
In classical SGD, solve recursively for $\theta_n$ to derive:

$$
\theta_n = \mathbf{P}_1^n \theta_0 + \sum_{i=1}^{n} \gamma_i \mathbf{P}_{i+1}^n Y_i X_i,
$$

where $\mathbf{P}_i^n = (I - \gamma_n X_n X_n^T) \cdots (I - \gamma_i X_i X_i^T)$.

$\mathbf{P}_1^n$ discounts initial conditions $\theta_0$. Its “size” is critical for stability:

$$(1 - \gamma_1 \lambda_j)(1 - \frac{\gamma_1}{2} \lambda_j) \cdots (1 - \frac{\gamma_1}{n} \lambda_j) \in \text{eig}(\mathbb{E}(\mathbf{P}_1^n)),
$$

where $\lambda_j$ is $j$th eigenvalue of $\mathbb{E}(XX^T)$.
In classical SGD, solve recursively for $\theta_n$ to derive:

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

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

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

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

where $\lambda_j$ is $j$th eigenvalue of $\mathbb{E}(XX^T)$.

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

$$\max_{n>0} \max \{\text{eig}(\mathbb{E}(P_1^n))\} = O(2^{\gamma_1 \lambda_{\text{max}} / \sqrt{\gamma_1 \lambda_{\text{max}}}}).$$
Quick recap

- SGD can be readily applied when $\nabla \log f$ is available.
- (Very) sensitive to specifications of learning rate – more on that later.
- Versatile tool for estimation – fast and relatively precise.
- How much precise?
Statistical efficiency of SGD procedures

Theorem (Toulis et. al., 2014)

Generalization | Nonasymptotics

Suppose $2\gamma_1 \mathcal{I}(\theta_\star) - \mathbb{I} \succ 0$. Then, the variance of SGD estimator satisfies:

$$n \text{Var}(\theta_n) \rightarrow \gamma_1^2 (2\gamma_1 \mathcal{I}(\theta_\star) - \mathbb{I})^{-1} \mathcal{I}(\theta_\star).$$
Suppose $2\gamma_1 I(\theta^\star) - \mathbb{I} \succ 0$. Then, the variance of SGD estimator satisfies:

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

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

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

- If $\gamma_1 < 1/(2\lambda_{\text{min}})$ then arbitrary inefficiency, e.g., $n^\epsilon \text{Var}(\theta_n) \to \infty$. 
Normal example (cont.)

Here, we plot $N(\theta_N - \theta^\star)^\top \Sigma^{-1}(\theta_N - \theta^\star)$ against $\chi^2_p$ distribution.

$\gamma_1 = 1$, $N = 2e4$, $p = 20$, and $\Sigma$ is the theoretical asymptotic variance of SGD.
Here, we plot $N(\theta_N - \theta_\star)^\top \Sigma^{-1}(\theta_N - \theta_\star)$ against $\chi^2_p$ distribution. 

$\gamma_1 = 10$, $N = 2e4$, $p = 20$, and $\Sigma$ is the theoretical asymptotic variance of SGD.
Efficiency loss of SGD

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

It follows, $\gamma_1^2 \lambda_j^2 \gamma_1 \lambda_j - 1 \in \text{eig}(\Sigma_{\theta^*, \gamma_1})$, 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 $\lambda_{\text{max}} > \lambda_{\text{min}}$ (spectral gap).
Efficiency loss of SGD

Let $\Sigma_{\theta_\star, \gamma_1} \triangleq \lim n \mathbb{V} \text{ar}(\theta_n) = \gamma_1^2 (2 \gamma_1 \mathcal{I}(\theta_\star) - \mathbb{I})^{-1} \mathcal{I}(\theta_\star)$.

It follows,

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

Implies efficiency loss because $\mathcal{I}(\theta_\star) - 1$ 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).
Efficiency loss of SGD

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

- It follows,

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

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

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

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

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

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

Efficiency loss of SGD

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

It follows,

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

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

- 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 \( \lambda_{\text{max}} >> \lambda_{\text{min}} \) (spectral gap).
Another principled way to set the optimal rate:

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

Need to estimate \( \lambda_j \); or terms can be approximated:

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

\[ \text{tr}(\Sigma_{\theta^*,\gamma_1}) \approx \gamma_1^2. \]

In fact, \( \Sigma_{\theta^*,\gamma_1} \approx \gamma_1^2 I \).
Another principled way to set the optimal rate:

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

Need to estimate \( \lambda_j \); or terms can be approximated:

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

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

- SGD can be readily applied when $\nabla \log f$ is available.
- Closed-form theoretical efficiency.
- Numerical stability: (very) sensitive to specifications of learning rate...
- Fix instability?
Fixing stability issues?

- Back to 2013 when I started working on such problems at Google.

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

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

- Back to 2013 when I started working on such problems at Google.

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

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

- Estimation through SGD uses discrepancies (observed \(n\) — expected \(n\)):

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

- Back to 2013 when I started working on such problems at Google.

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

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

- Estimation through SGD uses discrepancies (observed $n$ – 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 **implicit** update:

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

(See ⊳ intuition. See ⊳ computation.)
Estimation through SGD with **implicit** update:

\[
θ_n = θ_{n-1} + \frac{1}{n} (Y_n - e^{X_n'θ_n})X_n.
\]

(See > intuition. See > computation.)

**Example:** \(θ_0 = 0, Y_1 = Y_2 = 1001, X_n = 1\).

- **First iteration,**
  \[
  θ_1 = 0 + 1 \cdot (1001 - e^{θ_1}).
  \]
  Thus, \(θ_1 \approx \log(994) \approx 6.902\).

- **Second iteration,**
  \[
  θ_2 = 6.902 + \frac{1}{2} (1001 - e^{θ_2}).
  \]
  Thus, \(θ_2 \approx \log(1001) \approx 6.909\).
\[ \theta_\star = (0.47, 0.22, 0.1, \ldots)^\top \in \mathbb{R}^{20}, \quad \lambda_{\min} = \lambda_{\max} = 1, \quad \gamma_1 = 1. \]
\[ \theta_\star = (0.47, 0.22, 0.1, \ldots)^T \in \mathbb{R}^{20}, \quad \lambda_{\text{min}} = \lambda_{\text{max}} = 1, \quad \gamma_1 = 10. \]
Extension: parameter transformation

- No loss when $\mathcal{I}(\theta) \propto \mathbb{I}$, i.e., $\theta$ is location parameter.
- Suggests that transformations can be helpful.
Extension: parameter transformation

- No loss when $\mathcal{I}(\theta) \propto \mathbb{I}$, i.e., $\theta$ is location parameter.
- Suggests that transformations can be helpful.

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:

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

- No loss when $\mathcal{I}(\theta) \propto \mathbb{I}$, i.e., $\theta$ is **location parameter**.
- Suggests that transformations can be helpful.

**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, $\mathcal{I}(\mu) = \mathcal{I}(\theta)(\frac{d\theta}{d\mu})^2 = 1$. 
No loss when $\mathcal{I}(\theta) \propto \mathbb{I}$, i.e., $\theta$ is *location parameter*. 
Suggests that transformations can be helpful.

**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:

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

- Here, $\mathcal{I}(\theta) = 15\theta^4 \equiv \psi > 0$; but, $\mathcal{I}(\mu) = \mathcal{I}(\theta) (\frac{d\theta}{d\mu})^2 = 1$.
- For vanilla SGD, $n \text{Var}(\theta_n) \to \psi/(2\psi - 1)$. 
Extension: parameter transformation

- No loss when $\mathcal{I}(\theta) \propto \mathbb{I}$, i.e., $\theta$ is location parameter.
- Suggests that transformations can be helpful.

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, $\mathcal{I}(\mu) = \mathcal{I}(\theta) (\frac{d\theta}{d\mu})^2 = 1$.

- 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!
Extension: parameter transformation

- No loss when $\mathcal{I}(\theta) \propto I$, i.e., $\theta$ is location parameter.
- Suggests that transformations can be helpful.

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, $\mathcal{I}(\mu) = \mathcal{I}(\theta)(\frac{d\theta}{d\mu})^2 = 1$.
- For vanilla SGD, $n \mathbb{V} \mathbb{A}r(\theta_n) \rightarrow \psi/(2\psi - 1)$.
- For transformed, $n \mathbb{V} \mathbb{A}r(\mu_n) \rightarrow 1$ and $n \mathbb{V} \mathbb{A}r(\theta_n) \rightarrow 1/\psi = \mathcal{I}(\theta)^{-1} - \text{optimal!}$
- Unfortunately, general case is hard: find transformation with Jacobian $J$ s.t. $J\mathcal{I}(\theta)J^T \approx \text{const.}$
Parameter transformation – example

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

![Graph showing red = transformed; black = vanilla (implicit)]
An important extension is the following problem:

\[
\text{find } \theta_* \text{ s.t. } \mathbb{E} \left( \nabla \log f(Y; X, \psi_*) \right) = 0,
\]

where \( \psi = \Psi(\theta) \).

Here, \( \Psi(\theta) \) is hard to compute.

For example, in dynamic discrete choice models, \( \psi \) may denote the value function for different states/decisions., and \( \theta \) may denote the structural parameters.

Using SGD may be possible if we run two parallel procedures.
In many cases compute the likelihood may be prohibitive.

For example, let $f_\theta(G)$ be (proportionally) a network likelihood, and $S(G)$ be a vector of sufficient statistics; the normalizing constant is equal to $[\sum_G f_\theta(G)]^{-1}$.

Using SGD may be possible through the procedure:

$$\theta_n = \theta_{n-1} + \gamma_n (S^\text{sim}_n - S^\text{obs}_n).$$

Here $S^\text{sim}_n$ is simulated statistics conditional on $\theta_{n-1}$, e.g., through Metropolis-Hastings, which does not require the normalizing constant; $S^\text{obs}_n$ is the observed statistics.
Machine learning reviews


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 R.)
- PT, D Tran, EM Airoldi, ”Towards stability and optimality in stochastic gradient descent” (2016, AI & Statistics, AISTATS)
\[ \theta_\star = (0.37, 0.15)^\top, \quad \lambda_{\text{min}} = 1, \quad \lambda_{\text{max}} = 1, \quad \gamma_1 \in \{0.1, 1, 10\}. \]
\( \theta_* = (0.37, 0.15)^T, \lambda_{\text{min}} = 1, \lambda_{\text{max}} = 10, \gamma_1 \in \{0.1, 1, 10\} \).
Intuition: implicit update as an infinite series of standard updates:

\[ \theta(0)_{n+1} = \theta_n - \gamma_n (Y_n - e^{\theta_n}) \]

\[ \theta(1)_{n+1} = \theta_n - \gamma_n (Y_n - e^{\theta(0)_n}) \]

\[ \theta(2)_{n+1} = \theta_n - \gamma_n (Y_n - e^{\theta(1)_n}) \]

\[ \cdots \]

\[ \theta(\infty)_{n+1} = \theta_n - \gamma_n (Y_n - e^{\theta(\infty)_n}) \]

Intuition: implicit update as an infinite series of standard updates:

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

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

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

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

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

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

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

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

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

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

\[\cdots\]

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


Back to main.
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,

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

$$
= \theta_{n-1}^{\text{im}} + \gamma_n \xi_n (Y_n - e^{X_n^T\theta_{n-1}^{\text{im}}}) X_n \tag{3}
$$

$$
\triangleq \theta_{n-1}^{\text{im}} + a_n X_n. \tag{4}
$$

Equate the two scales:

$$
a_n = \gamma_n (Y_n - e^{X_n^T\theta_{n}^{\text{im}}}) \text{[by setting (1) = (3)]} = \gamma_n (Y_n - e^{X_n^T\theta_{n-1}^{\text{im}} - 1 + ||X_n||_2^2}) a_n \text{[by substituting } \theta_{n}^{\text{im}} \text{ with (3)]}
$$

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

$$
x = a - bx^c,
$$

where $b,c > 0$. It follows that $x \in [\min(0, a - b), \max(0, a - b)]$. Back to main.
Efficient computation of implicit updates

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

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

$$\Rightarrow \theta_{im}^{n} = \theta_{im}^{n-1} + \gamma_n \xi_n (Y_n - e^{X_n^T\theta_{im}^{n-1}})X_n$$  \hspace{1cm} (3)

$$\triangleq \theta_{im}^{n-1} + a_n X_n.$$  \hspace{1cm} (4)

Equate the two scales:

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

$$= \gamma_n (Y_n - e^{X_n^T\theta_{im}^{n-1}} + \|X_n\|^2 a_n).$$  \hspace{1cm} [by substituting $\theta_{im}^{n}$ 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.
**Example.** Estimate CDF $F(t)$ with data $Y_1, Y_2, \ldots, Y_n$; $Y^\text{obs} = \text{uncensored}$. 
**Self-consistency principle**

- **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

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

Back to [main >](#).
In an experiment, suppose $\theta$ is input, $H(\theta)$ random output. Suppose we wish to find $\theta^\star$ such that

$$\mathbb{E}(H(\theta^\star)) = 0.$$
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): $\mathbb{E}\left(\left|\theta_n - \theta^\star\right|^2\right) \to 0$ if
- $\sum \gamma_i = \infty$; $\sum \gamma_i^2 < \infty$;
- $H$ is concave in expectation and Lipschitz;
- $\mathbb{E}\left(||H(\theta^\star)||^2\right) < \infty$.

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)
  \[
  \theta_n = \theta_{n-1} + \gamma_n H(\theta^*_{n-1})
  \]
  \[
  \text{s.t. } E(\theta_n|\theta_{n-1}) = \theta^*_{n-1}
  \]

- Non-asymptotic/asymptotic analysis (Toulis & Airoldi, 2015b)
- Implementations need to estimate $\theta^*_{n-1}$
Optimal efficiency: second-order SGD

**Theorem (Toulis & Airoldi, 2015a)**

Consider the second-order implicit SGD procedure

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

where \( C_n \to C > 0 \), where \( C \) is symmetric and commutes with \( I(\theta^*) \).

Then

\[
n \nabla \text{Var}(\theta_n^{im}) \to (2C I(\theta^*) - I)^{-1} C I(\theta^*) C \equiv \Sigma_{\theta^*,C}.
\]
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 \( \mathcal{I}(\theta_*) \).

Then

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

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

Theorem (Toulis & Airoldi, 2015a)

Consider the second-order implicit SGD procedure

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

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

Then

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

- Optimal efficiency **only** if \( C = \mathcal{I}(\theta_*)^{-1} \).
- **Adaptive** methods concurrently estimate \( \mathcal{I}(\theta_*)^{-1} \);
  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.

Panagiotis (Panos) Toulis
Review of stochastic optimization 35 / 27
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 \( C_n \rightarrow \text{diag}(\mathcal{I}(\theta_*)^{-1}) \).
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)

\[ \sqrt{n} \text{Var}(\theta_{n}^{\text{ada}}) \to \frac{\gamma_{1}}{2} \text{diag}(\mathcal{I}(\theta^*))^{-1/2}. \] \hspace{1cm} (5)

- AdaGrad is inefficient but (1) holds \textbf{regardless} of \( \gamma_{1} \).
- In contrast, SGD procedures require \( \gamma_{1} > 1/(2\lambda_{\text{min}}) \) for \( O(1/n) \) efficiency.
\[ \theta_* = (2.23, 0.5, 0.1, 0.02, 0.01)^\top; \quad \lambda_j \in [1, 10] \]
\[ \theta_n = \theta_{n-1} + \gamma_n H(\theta^*_{n-1}) \]
\[
\text{s.t. } \mathbb{E}(\theta_n | \theta_{n-1}) = \theta^*_{n-1}
\]

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

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

- \( x_k \to \theta^*_{n-1} \) (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 \( \gamma_n \propto n^{-\gamma} \), \( \gamma \in (0, 1) \), \( \lambda_{\text{min}} > 0 \),

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

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

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

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

- \( \lambda_{\text{min}} > 0 \) 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.
Implicit SGD can be written as

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

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

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

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

- 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)
In optimization problem, \( \arg \min_{\theta} g(\theta) \), for deterministic \( g \) we can do

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

- 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):
  \[
  \theta_n = \text{prox}_{\gamma_n R} \left( \theta_{n-1} + \gamma_n \nabla \log f(Y_n; X_n, \theta_{n-1}) \right)
  \]
  - \( R \) 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=\#\text{datapoints}, i=\text{datapoint index}, f_i=\text{loss at } i \text{ 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.
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 $\gamma_1 >> 1/(2\lambda_{\text{min}})$,

$$\text{tr}(\Sigma_{\theta^*,\gamma_1}) \approx p \frac{\gamma_1}{2}. \text{ In fact, } \Sigma_{\theta^*,\gamma_1} \approx \frac{\gamma_1}{2} I \text{ (parameter-free!) }$$

Fairly general way to construct pivotal quantity for $\theta_*$.

But we pay price in efficiency.

Back to optimal rates.
Standard asymptotic analysis obtains recursion for $\mathbb{E} \left( ||\theta^e_n - \theta^*||^2 \right)$. However, in the implicit procedure $\theta_{im} = \theta_{im} - \gamma_n \nabla \log f(Y_n; X_n, \theta_{im})$, we cannot use standard analysis because $(Y_n, X_n) \not\perp \perp \theta_{im}$.
Standard asymptotic analysis obtains recursion for $\mathbb{E} \left( \| \theta_n^{\text{ex}} - \theta_* \|^2 \right)$. 

A crucial property is the concavity of 

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

which requires 

$$(Y_n, X_n) \perp \perp \theta_{n-1}^{\text{ex}}.$$
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

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

which requires

$$(Y_n, X_n) \perp \perp \theta_{n-1}^{\text{ex}}.$$  

- 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,

\[
\begin{align*}
\theta^\text{im}_n &= \theta^\text{im}_{n-1} + \gamma_n \nabla \log f(Y_n; X_n, \theta^\text{im}_n) \\
&= \theta^\text{im}_{n-1} + \gamma_n \xi_n \nabla \log f(Y_n; X_n, \theta^\text{im}_{n-1}).
\end{align*}
\]
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,

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

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

\( (Y_n, X_n) \perp \perp \theta^\text{im}_{n-1}. \)

Proceed with \( \text{analysis} \). Back to \( \text{main} \).
Almost-sure bound for $\xi_n$

- Start with

$$\theta_{n}^{im} = \theta_{n-1}^{im} + \gamma_n \xi_n \nabla \log f(Y_n; X_n, \theta_{n-1}^{im}).$$
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 $\hat{\mathcal{I}}(\theta) = -\nabla^2 \log f(Y; X, \theta)$ and suppose $\text{tr}(\hat{\mathcal{I}}(\theta)) \geq s > 0$.

- 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,
  \[
  \mathbb{E} (||\theta_{n}^{im} - \theta_{*}||^2) \leq \frac{1}{1 + \gamma_{n}s} \mathbb{E} (||\theta_{n-1}^{im} - \theta_{*}||^2) + O(\gamma_{n}^2).
  \]

Back to main. Proceed to solving the recursion.
Suppose we wish to solve $b_n \leq F(b_{n-1})$, $F$ non-decreasing.

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

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

(minorize) Minimize $c_n^*$ wrt $\alpha$ to min. upper bound, $b_n \leq c_n^*$. 

Panagiotis (Panos) Toulis
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 $\alpha_0 n^2 + \alpha_1 n$. 

Back to main ⊿.
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 \( c^\alpha_n = \alpha_0 n^2 + \alpha_1 n \) and solve:

\[
 c^\alpha_n \geq c^\alpha_{n-1} + n
\]

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

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

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

\[
 b_n \leq c^*_n = \arg \min_{\alpha} c^\alpha_n = .5n^2 + .5n = n(n + 1)/2
\]
Intractable likelihoods: Monte-Carlo SGD

- In many cases the likelihood is intractable, thus SGD cannot be used.
Intractable likelihoods: Monte-Carlo SGD

- 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), $T(\theta)$ = simulated average statistics.

By SA theory $\theta_n$ converges to point $\theta_\infty$ such that $T(\theta_\infty) = S^{obs}$.
Intractable likelihoods: Monte-Carlo SGD

- 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

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

Back to main.