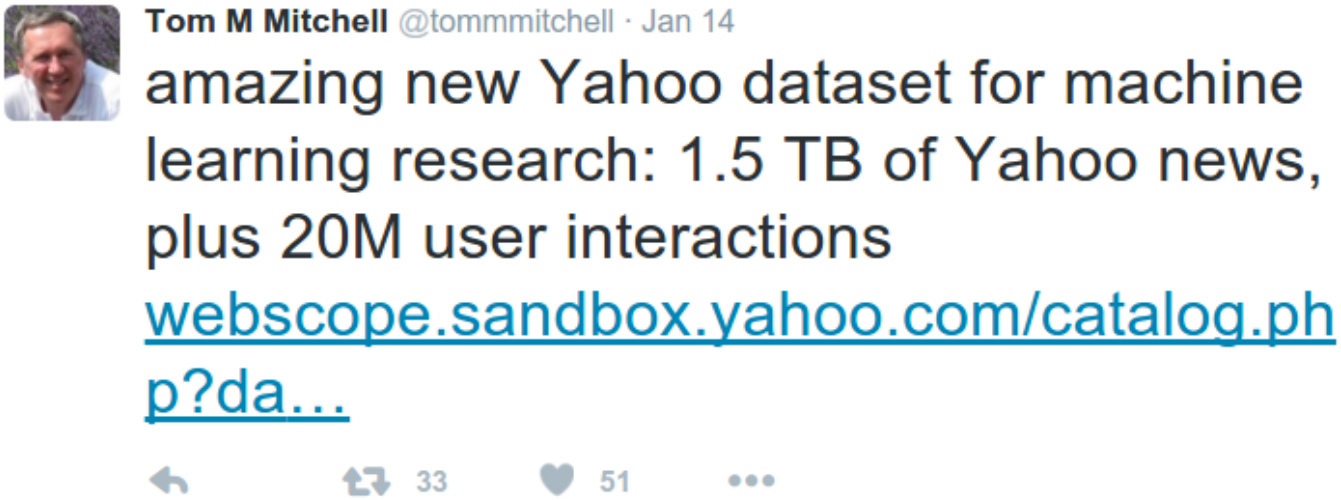
Asymptotic and finite-sample properties of estimators based on stochastic gradients

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

Econometrics and Statistics
University of Chicago, Booth School of Business



- Optimization and estimation are complementary.

Tom M Mitchell @tommmitchell · Jan 14

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...

33 51

- 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).

SGD in optimization and statistics

- 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 θ at ith datapoint.

SGD in optimization and statistics

- 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 θ at ith 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 in optimization and statistics

- 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 θ at ith 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 $[1, 2, \ldots, N]$.

SGD in optimization and statistics

- 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 θ at ith 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:

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

  where l_i is log-likelihood of θ at ith 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

- $Y = \#\text{bookings}$; $X = \text{covariates}$; $Y \sim \text{Poisson}(e^{X'\theta_\star})$.
- **Goal:** use i.i.d. data $(X_1, Y_1), (X_2, Y_2), \ldots,$ to estimate $\theta_\star$.

Motivation: modeling flight ticket booking

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

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

Motivation: modeling flight ticket booking

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

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

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 **implicit** update:

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

(See ▷ intuition. See ▷ computation.)

- Estimation through SGD with **implicit** update:

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

(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);
See incremental proximal method	(Bertsekas, 2011);
See 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);
See incremental proximal method	(Bertsekas, 2011);
See 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. We build towards doing 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. $\theta_\star \in \mathbb{R}^p$:

$$Y|X \sim f(Y; X, \theta_\star).$$

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

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

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

Asymptotic variance		Statistical efficiency
	$= \mathcal{I}(\theta_\star)^{-1}$	optimal efficiency;
$\lim_{n\to\infty} n\mathbb{V}\text{ar}(\hat{\theta}_n)$	$\succ \mathcal{I}(\theta_\star)^{-1}$	efficiency loss;
	$= \infty$	inefficiency.

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

- 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.

- 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:

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

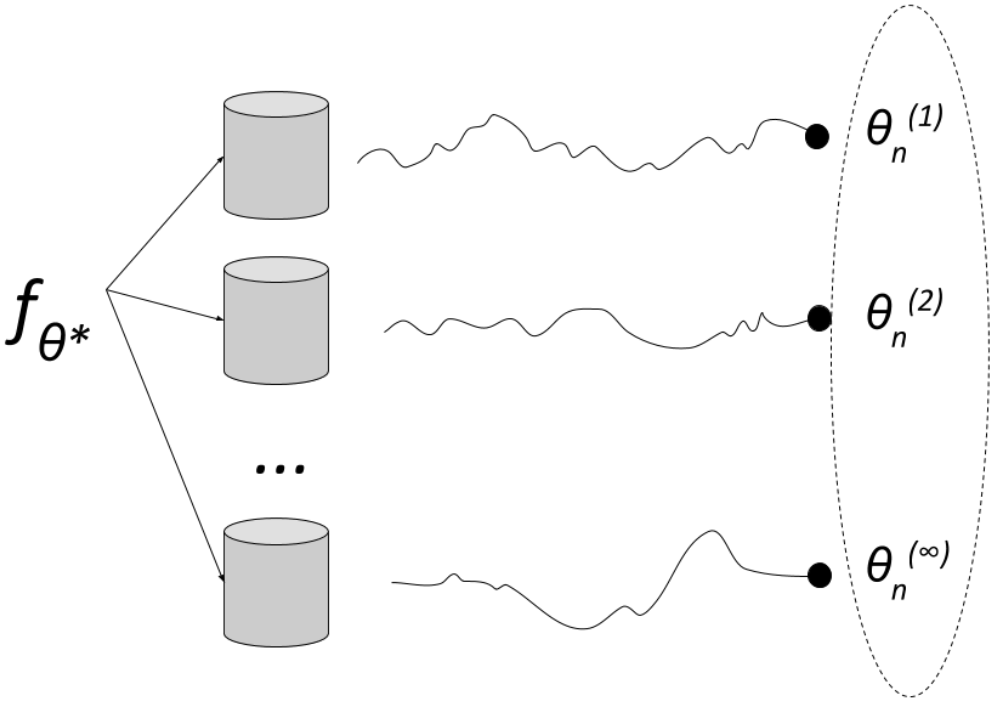
$$\boldsymbol{\theta_n^{\text{im}}} = \theta_{n-1}^{\text{im}} + \gamma_n \nabla \log f(Y_n; X_n, \boldsymbol{\theta_n^{\text{im}}}).$$ Bayesian interpretation ▷

Stochastic approximation theory (for explicit procedures only):

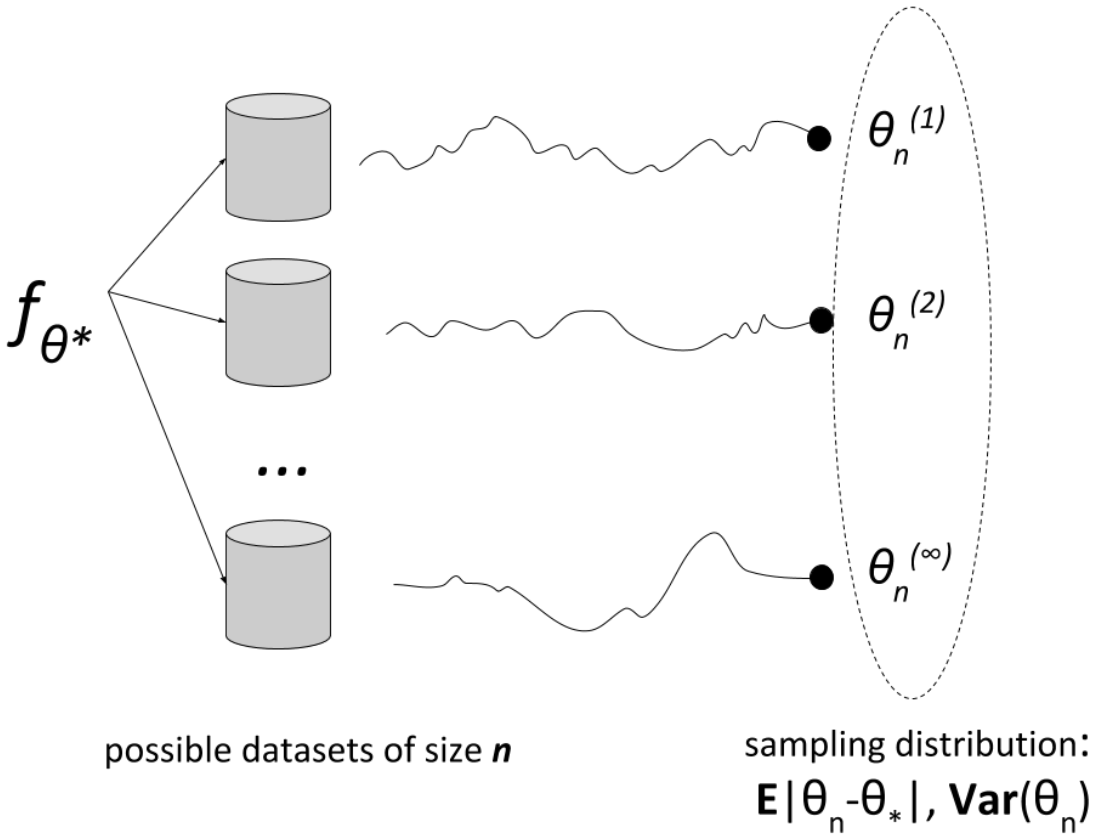
$$\mathbb{E}\left(\nabla \log f(Y; X, \theta_\infty)\right) = 0 \Leftrightarrow \theta_\infty = \theta_\star.$$ details ▷

Ideal for "big data" because likelihood is computed at single datapoint.

Focus on sampling variability



Focus on sampling variability



This talk focuses on two main points:

- Implicit SGD has better numerical stability without sacrificing efficiency;
- such stability enables statistical inference with large datasets.

Numerical stability: illustration on normal model

- Suppose normal model $Y|X \sim \mathcal{N}(X^\intercal \theta_\star, 1)$. Then,

$$\nabla \log f(Y; X, \theta) = (Y - X^\intercal \theta) X.$$

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

Numerical stability: illustration on normal model

- Suppose normal model $Y|X \sim \mathcal{N}(X^\intercal \theta_\star, 1)$. Then,

$$\nabla \log f(Y; X, \theta) = (Y - X^\intercal \theta) X.$$

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

- Explicit SGD estimator:

$$\begin{aligned}\theta_n^{\mathsf{ex}} &= \theta_{n-1}^{\mathsf{ex}} + \gamma_n (Y_n - X_n^\intercal \theta_{n-1}^{\mathsf{ex}}) X_n \\ &= (\mathbb{I} - \gamma_n X_n X_n^\intercal)\theta_{n-1}^{\mathsf{ex}} + \gamma_n Y_n X_n.\end{aligned}$$

Numerical stability: illustration on normal model

- Suppose normal model $Y|X \sim \mathcal{N}(X^\intercal \theta_\star, 1)$. Then,

$$\nabla \log f(Y; X, \theta) = (Y - X^\intercal \theta) X.$$

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

- Explicit SGD estimator:

$$\begin{aligned}\theta_n^{\mathsf{ex}} &= \theta_{n-1}^{\mathsf{ex}} + \gamma_n (Y_n - X_n^\intercal \theta_{n-1}^{\mathsf{ex}}) X_n \\ &= (\mathbb{I} - \gamma_n X_n X_n^\intercal)\theta_{n-1}^{\mathsf{ex}} + \gamma_n Y_n X_n.\end{aligned}$$

- Implicit SGD estimator:

$$\begin{aligned}\boldsymbol{\theta_n^{\mathsf{im}}} &= \theta_{n-1}^{\mathsf{im}} + \gamma_n (Y_n - X_n^\intercal \boldsymbol{\theta_n^{\mathsf{im}}}) X_n \\ &= (\mathbb{I} + \gamma_n X_n X_n^\intercal)^{-1} \theta_{n-1}^{\mathsf{im}} + \gamma_n (\mathbb{I} + \gamma_n X_n X_n^\intercal)^{-1} Y_n X_n.\end{aligned}$$

- In explicit SGD, solve recursively for θ_n^{ex} to derive:

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

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

- In explicit SGD, solve recursively for θ_n^{ex} to derive:

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

  where $\mathbf{P}_i^n = (\mathbb{I} - \gamma_n X_n X_n^\intercal) \cdots (\mathbb{I} - \gamma_i X_i X_i^\intercal)$.
- $\mathbf{P}_1^n$ discounts initial conditions θ_0^{ex}. Its "size" is critical for stability:

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

- In explicit SGD, solve recursively for θ_n^{ex} to derive:

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

  where $\mathbf{P}_i^n = (\mathbb{I} - \gamma_n X_n X_n^\intercal) \cdots (\mathbb{I} - \gamma_i X_i X_i^\intercal)$.
- $\mathbf{P}_1^n$ discounts initial conditions θ_0^{ex}. Its "size" is critical for stability:

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

- For stability it is desirable that

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

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

$$\max_{n>0} \max\{\mathsf{eig}(\mathbb{E}\,(\mathbf{P}_1^n))\} = O(2^{\gamma_1 \lambda_{\mathsf{max}}}/\sqrt{\gamma_1 \lambda_{\mathsf{max}}}).$$

- In implicit SGD, solve recursively for θ_n^{im} to derive:

$$\theta_n^{\text{im}} = \mathbf{Q}_1^n \theta_0^{\text{im}} + \sum_{i=1}^{n} \gamma_i \mathbf{Q}_{i+1}^n \mathbf{Q}_i^i Y_i X_i.,$$

where $\mathbf{Q}_i^n = (\mathbb{I} + \gamma_n X_n X_n^\intercal)^{-1} \cdots (\mathbb{I} + \gamma_i X_i X_i^\intercal)^{-1}$.

- In implicit SGD, solve recursively for θ_n^{im} to derive:

$$\theta_n^{\text{im}} = \mathbf{Q}_1^n \theta_0^{\text{im}} + \sum_{i=1}^{n} \gamma_i \mathbf{Q}_{i+1}^n \mathbf{Q}_i^i Y_i X_i.,$$

  where $\mathbf{Q}_i^n = (\mathbb{I} + \gamma_n X_n X_n^\intercal)^{-1} \cdots (\mathbb{I} + \gamma_i X_i X_i^\intercal)^{-1}$.
- $\mathbf{Q}_1^n$ discounts initial θ_0^{im}. Its "size" is critical for stability:

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

- In implicit SGD, solve recursively for θ_n^{im} to derive:

$$\theta_n^{\text{im}} = \mathbf{Q}_1^n \theta_0^{\text{im}} + \sum_{i=1}^{n} \gamma_i \mathbf{Q}_{i+1}^n \mathbf{Q}_i^i Y_i X_i.,$$

  where $\mathbf{Q}_i^n = (\mathbb{I} + \gamma_n X_n X_n^\intercal)^{-1} \cdots (\mathbb{I} + \gamma_i X_i X_i^\intercal)^{-1}$.
- $\mathbf{Q}_1^n$ discounts initial θ_0^{im}. Its "size" is critical for stability:

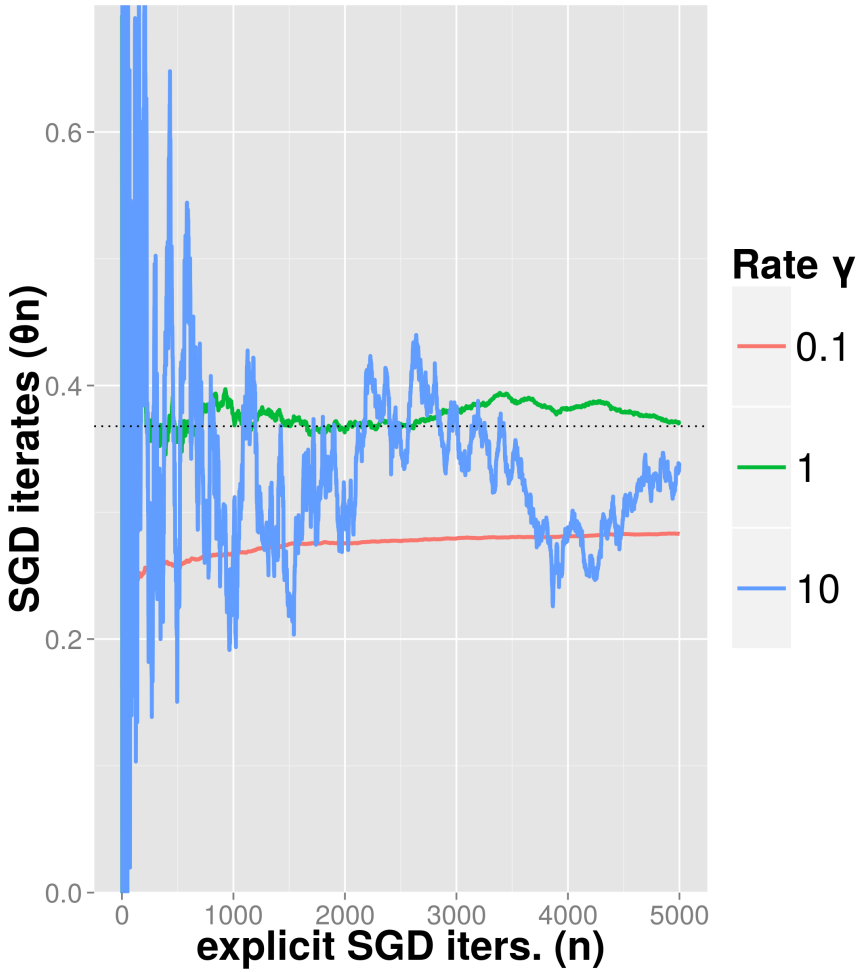
$$(1+\gamma_1\lambda_j)^{-1}(1+\frac{\gamma_1}{2}\lambda_j)^{-1}\cdots(1+\frac{\gamma_1}{n}\lambda_j)^{-1} \in \text{eig}(\mathbb{E}\,(\mathbf{Q}_1^n)).$$

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

$$\max_{n>0} \max\{\text{eig}(\mathbb{E}\,(\mathbf{Q}_1^n))\} = O(1).$$

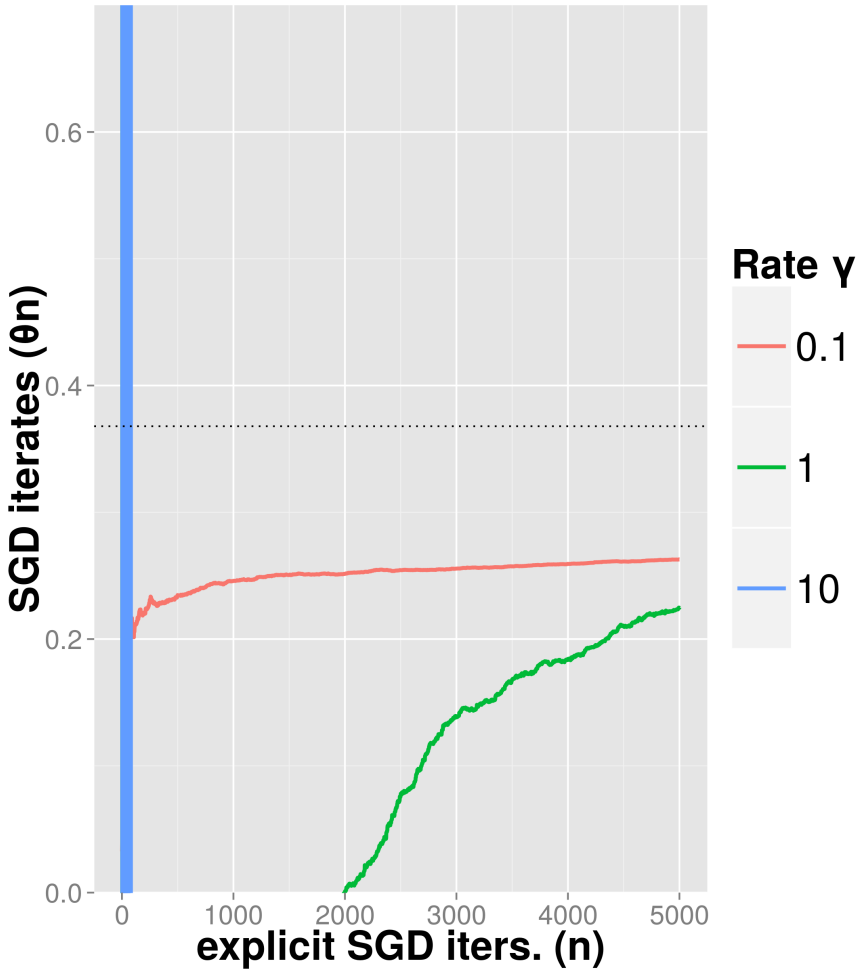
Stability simulation: easy case

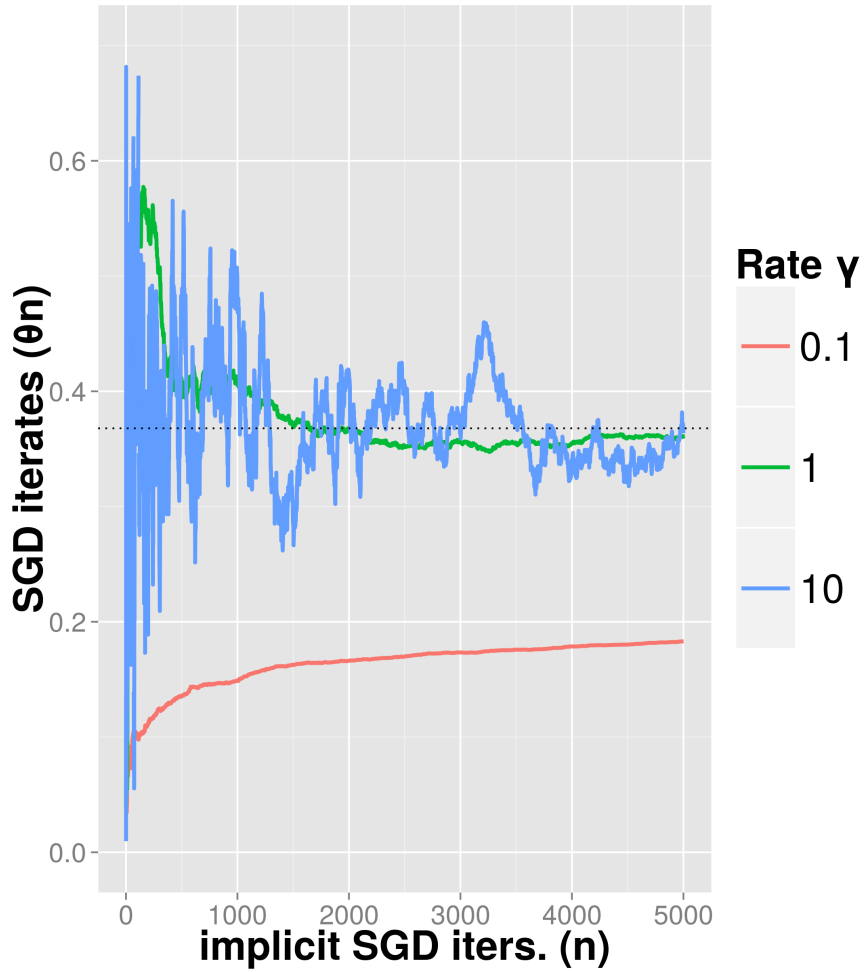
- $\theta_\star = (0.37, 0.15)^\intercal$, $\lambda_{\text{min}} = 1, \lambda_{\text{max}} = 1, \gamma_1 \in \{0.1, 1, 10\}$.



Stability simulation: divergence of explicit SGD

- $\theta_\star = (0.37, 0.15)^\intercal$, $\lambda_{\text{min}} = 1, \lambda_{\text{max}} = 10, \gamma_1 \in \{0.1, 1, 10\}$.





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 implicit SGD satisfies*

$$n\mathbb{V}\mathsf{ar}(\theta_n^{\mathsf{im}}) \to \gamma_1^2(2\gamma_1\mathcal{I}(\theta_\star) - \mathbb{I})^{-1}\mathcal{I}(\theta_\star).$$

The explicit SGD estimator has the **same** *asymptotic efficiency.*

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 implicit SGD satisfies*

$$n\mathbb{V}\mathsf{ar}(\theta_n^{\mathsf{im}}) \to \gamma_1^2(2\gamma_1\mathcal{I}(\theta_\star) - \mathbb{I})^{-1}\mathcal{I}(\theta_\star).$$

The explicit SGD estimator has the **same** *asymptotic efficiency.*

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

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

- If $\gamma_1 < 1/(2\lambda_{\mathsf{min}})$ then arbitrary inefficiency, e.g., $n^\epsilon\mathbb{V}\mathsf{ar}(\theta_n^{\mathsf{im}}) \to \infty$.

Summarizing the constraints

	Explicit SGD	Implicit SGD
Stability	$\gamma_1 < 2/\lambda_{\mathsf{max}}$	-
Efficiency	$\gamma_1 > 1/(2\lambda_{\mathsf{min}})$	$\gamma_1 > 1/(2\lambda_{\mathsf{min}})$

Summarizing the constraints

	Explicit SGD	Implicit SGD
Stability	$\gamma_1 < 2/\lambda_{\text{max}}$	-
Efficiency	$\gamma_1 > 1/(2\lambda_{\text{min}})$	$\gamma_1 > 1/(2\lambda_{\text{min}})$

- The requirements for explicit SGD are very 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)

Summarizing the constraints

	Explicit SGD	Implicit SGD
Stability	$\gamma_1 < 2/\lambda_{\text{max}}$	-
Efficiency	$\gamma_1 > 1/(2\lambda_{\text{min}})$	$\gamma_1 > 1/(2\lambda_{\text{min}})$

- The requirements for explicit SGD are very 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)

Two main talk points:

- ✓ Implicit SGD has better numerical stability without sacrificing efficiency;
- such stability enables principled statistical analysis with large datasets.

Efficiency loss of SGD

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

Efficiency loss of SGD

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

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

Efficiency loss of SGD

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

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

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

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

Efficiency loss of SGD

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

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

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

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

Efficiency loss of SGD

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

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

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

- 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 $\lambda_{\text{max}} >> \lambda_{\text{min}}$ (spectral gap).

Optimal rates

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

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

Optimal rates

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

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

- Note $\frac{\gamma_1^2 \lambda_j}{2\gamma_1 \lambda_j - 1} \approx \frac{\gamma_1}{2}$ if $2\gamma_1 \lambda_j >> 1$.

Optimal rates

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

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

- 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 $\gamma_1 > 1/(2\lambda_{\mathsf{min}})$,

$$\mathrm{tr}(\Sigma_{\theta_\star,\gamma_1}) \approx (p-1)\frac{\gamma_1}{2} + \frac{\gamma_1^2 \lambda_{\mathsf{min}}}{2\gamma_1 \lambda_{\mathsf{min}} - 1}.$$

Optimal rates

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

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

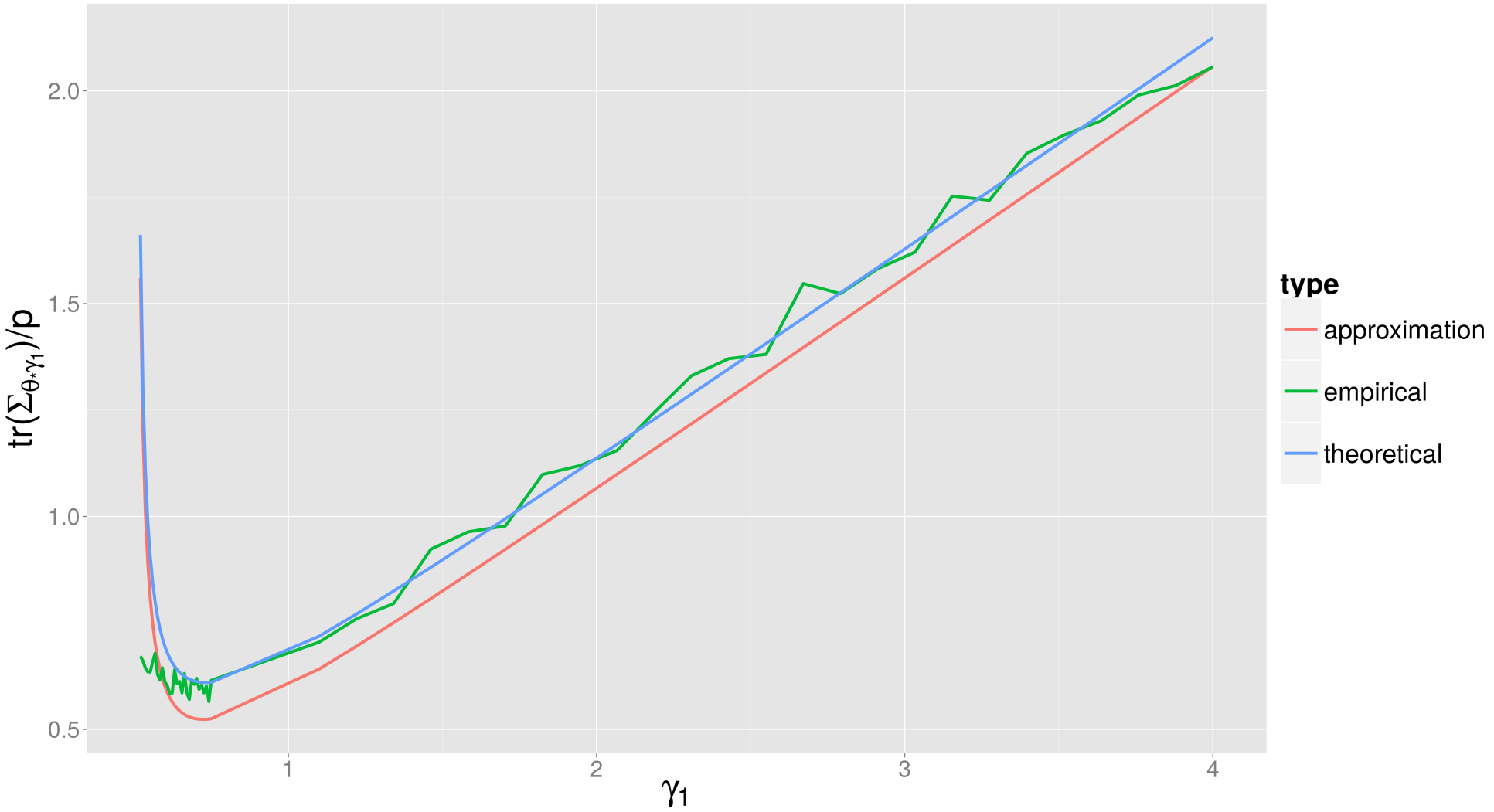
- Note $\frac{\gamma_1^2 \lambda_j}{2\gamma_1 \lambda_j - 1} \approx \frac{\gamma_1}{2}$ if $2\gamma_1 \lambda_j >> 1$.
- If $\gamma_1 > 1/(2\lambda_{\mathsf{min}})$,

$$\mathrm{tr}(\Sigma_{\theta_\star,\gamma_1}) \approx (p-1)\frac{\gamma_1}{2} + \frac{\gamma_1^2 \lambda_{\mathsf{min}}}{2\gamma_1 \lambda_{\mathsf{min}} - 1}.$$

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

$$\mathrm{tr}(\Sigma_{\theta_\star,\gamma_1}) \approx p\frac{\gamma_1}{2} \quad (\text{In fact, } \Sigma_{\theta_\star,\gamma_1} \approx \frac{\gamma_1}{2}\mathbb{I}).$$ pivotal quantity ▷

- Normal model, $\lambda_j \in \{1, 2, \ldots 5\}$, need $\gamma_1 > 1/(2\lambda_{\text{min}}) = 0.5$.



Asymptotic normality

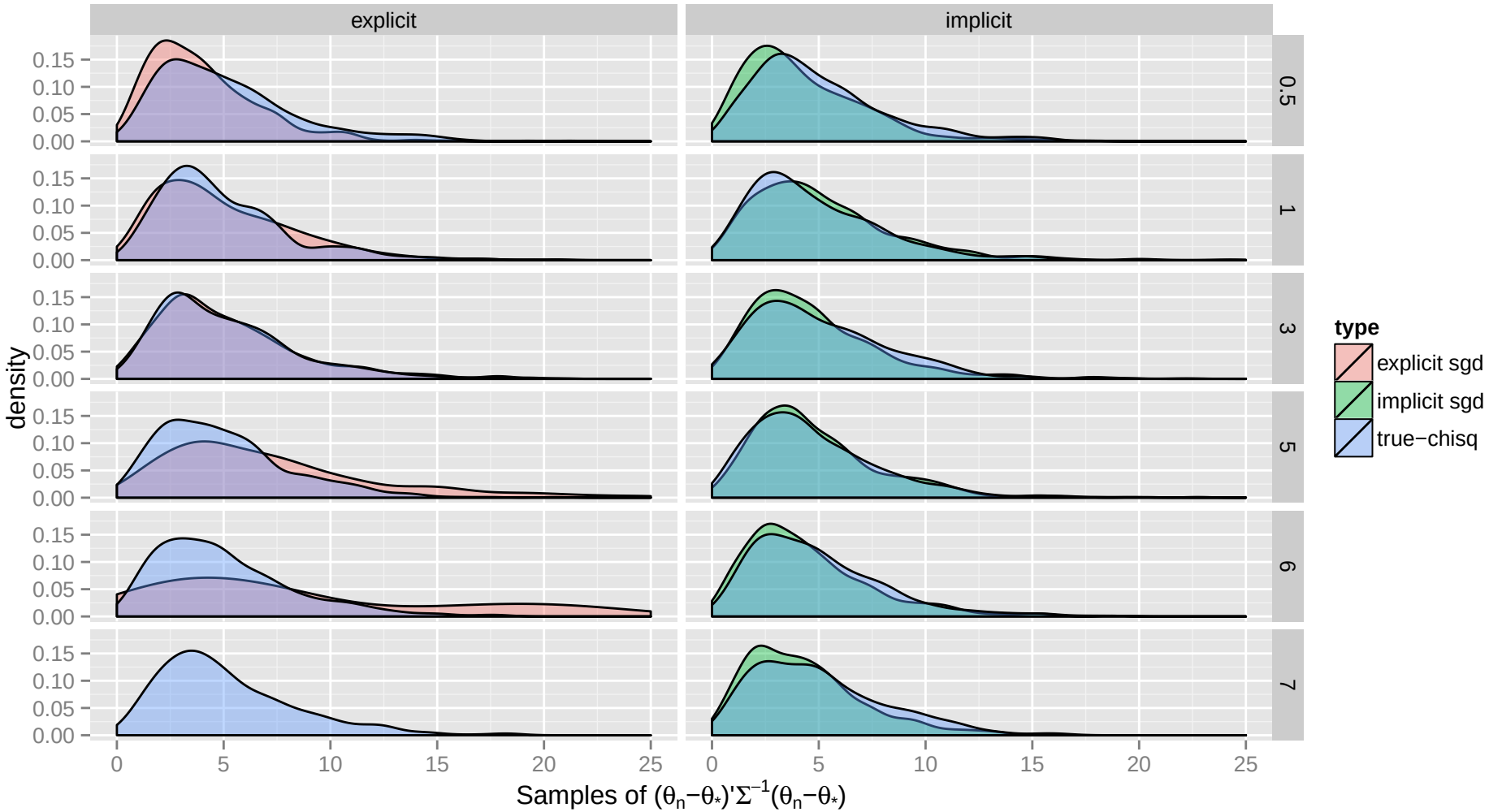
- Under typical Lindeberg conditions,

$$\sqrt{n}(\theta_n^{\mathsf{im}} - \theta_\star) \to \mathcal{N}(0, \Sigma_{\theta_\star, \gamma_1}).$$

Asymptotic normality

- Under typical Lindeberg conditions,

$$\sqrt{n}(\theta_n^{\text{im}} - \theta_\star) \to \mathcal{N}(0, \Sigma_{\theta_\star, \gamma_1}).$$



Summing up the good properties of implicit SGD

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:

- ✓ Implicit SGD has better numerical stability without sacrificing efficiency;
- ✓ such stability enables principled statistical analysis with large datasets.

Ongoing work

- Statistics of optimization procedures (e.g., Second-order ▷ procedures).
- Network models/intractable likelihoods (e.g., Monte-Carlo SGD ▷).
- Combo: search with constant rate, then converge with decreasing rate.
- Reinforcement learning and neural networks.

THANK YOU!

References

- PT, EM Airoldi, "Asymptotic and finite-sample properties of estimators based on stochastic gradients." (2016, Annals of Stat., in press)
- PT, EM Airoldi, "Scalable estimation strategies based on stochastic approximations: classical results and new insights" (2015, Statistics and Computing)
- PT, J Rennie, EM Airoldi, "Statistical analysis of stochastic gradient methods for generalized linear models", (2014, Int'l Conference in Machine Learning, ICML)
- PT, D Tran, EM Airoldi, "Towards stability and optimality in stochastic gradient descent" (2016, AI & Statistics, AISTATS)
- D Tran, PT, Edoardo M. Airoldi, "sgd R package: stochastic gradient methods for estimation with large data sets" (2016, in review)

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

- 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 **infinite** series of standard updates:

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

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

- 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 **infinite** series of standard updates:

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

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

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

$$\dots$$

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

- cf. self-consistency ▷ principle in statistics (Efron, 1967); (Tarpey & Flury, 1996).
- Back to main.

Efficient computation of implicit updates

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

Efficient computation of implicit updates

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

$$\begin{aligned}
\theta_n^{\text{im}} &= \theta_{n-1}^{\text{im}} + \gamma_n (Y_n - e^{X_n^\intercal \theta_n^{\text{im}}}) X_n && (1)\\
&= \theta_{n-1}^{\text{im}} + \gamma_n \xi_n (Y_n - e^{X_n^\intercal \theta_{n-1}^{\text{im}}}) X_n && (2)\\
&\triangleq \theta_{n-1}^{\text{im}} + a_n X_n. && (3)
\end{aligned}$$

Efficient computation of implicit updates

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

$$\begin{aligned}
\theta_n^{\text{im}} &= \theta_{n-1}^{\text{im}} + \gamma_n (Y_n - e^{X_n^\intercal \theta_n^{\text{im}}}) X_n && (1)\\
&= \theta_{n-1}^{\text{im}} + \gamma_n \xi_n (Y_n - e^{X_n^\intercal \theta_{n-1}^{\text{im}}}) X_n && (2)\\
&\triangleq \theta_{n-1}^{\text{im}} + a_n X_n. && (3)
\end{aligned}$$

Equate the two scales:

$$\begin{aligned}
a_n &= \gamma_n (Y_n - e^{X_n^\intercal \theta_n^{\text{im}}}) && \text{[by setting (1) = (3)]}\\
&= \gamma_n (Y_n - e^{X_n^\intercal \theta_{n-1}^{\text{im}} + \|X_n\|^2 a_n}). && \text{[by substituting } \theta_n^{\text{im}} \text{ with (3)]}
\end{aligned}$$

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

$$x = a - b e^{cx},$$

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 $Y_1, Y_2, \ldots, Y_n$; $\boldsymbol{Y}^{\text{obs}} =$ uncensored.

Self-consistency principle

- **Example.** Estimate CDF $F(t)$ with data $Y_1, Y_2, \ldots, Y_n$; $\boldsymbol{Y}^{\text{obs}}$= uncensored.
- A self-consistent estimator of $F(t)$ is

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

Back to main ▷.

Stochastic approximation

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

$$\mathbb{E}\left(H(\theta_\star)\right) = 0.$$

Stochastic approximation

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

$$\mathbb{E}\left(H(\theta_\star)\right) = 0.$$

- Robbins-Monro (1951) stochastic approximation procedure:

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

- Theorem (Robbins and Monro, 1951): $\mathbb{E}\left(|\theta_n - \theta_\star|^2\right) \to 0$ if
  - $\sum \gamma_i = \infty$; $\sum_i \gamma_i^2 < \infty$;
  - 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

$$\mathbb{E}\left(\nabla \log f(Y; X, \theta_\star)\right) = 0.$$

Back to main ▷.

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)

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

- Non-asymptotic/asymptotic analysis (Toulis & Airoldi, 2015b)
- Implementations need to estimate θ^*_{n-1}

Optimal efficiency: second-order SGD

Theorem (Toulis & Airoldi, 2015a)

Consider the second-order implicit SGD procedure

$$\theta_n^{\mathsf{im}} = \theta_{n-1}^{\mathsf{im}} + \frac{1}{n}\boldsymbol{C}_n \nabla \log f(Y_n; X_n, \theta_n^{\mathsf{im}}),$$

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

$$n\mathbb{V}\mathsf{ar}(\theta_n^{\mathsf{im}}) \to (2C\mathcal{I}(\theta_\star) - \mathbb{I})^{-1} C\mathcal{I}(\theta_\star)C \triangleq \Sigma_{\theta_\star, C}.$$

Optimal efficiency: second-order SGD

Theorem (Toulis & Airoldi, 2015a)

Consider the second-order implicit SGD procedure

$$\theta_n^{\mathsf{im}} = \theta_{n-1}^{\mathsf{im}} + \frac{1}{n}\boldsymbol{C}_n \nabla \log f(Y_n; X_n, \theta_n^{\mathsf{im}}),$$

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

$$n\mathbb{V}\mathsf{ar}(\theta_n^{\mathsf{im}}) \to (2C\mathcal{I}(\theta_\star) - \mathbb{I})^{-1} C\mathcal{I}(\theta_\star)C \triangleq \Sigma_{\theta_\star, C}.$$

- 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

$$\theta_n^{\mathsf{im}} = \theta_{n-1}^{\mathsf{im}} + \frac{1}{n} \boldsymbol{C}_n \nabla \log f(Y_n; X_n, \theta_n^{\mathsf{im}}),$$

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

$$n\mathbb{V}\mathsf{ar}(\theta_n^{\mathsf{im}}) \to (2C\mathcal{I}(\theta_\star) - \mathbb{I})^{-1} C\mathcal{I}(\theta_\star)C \triangleq \Sigma_{\theta_\star, C}.$$

- 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 $C_n \to \text{diag}(\mathcal{I}(\theta_\star)^{-1})$.

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 $C_n \to \text{diag}(\mathcal{I}(\theta_\star)^{-1})$.

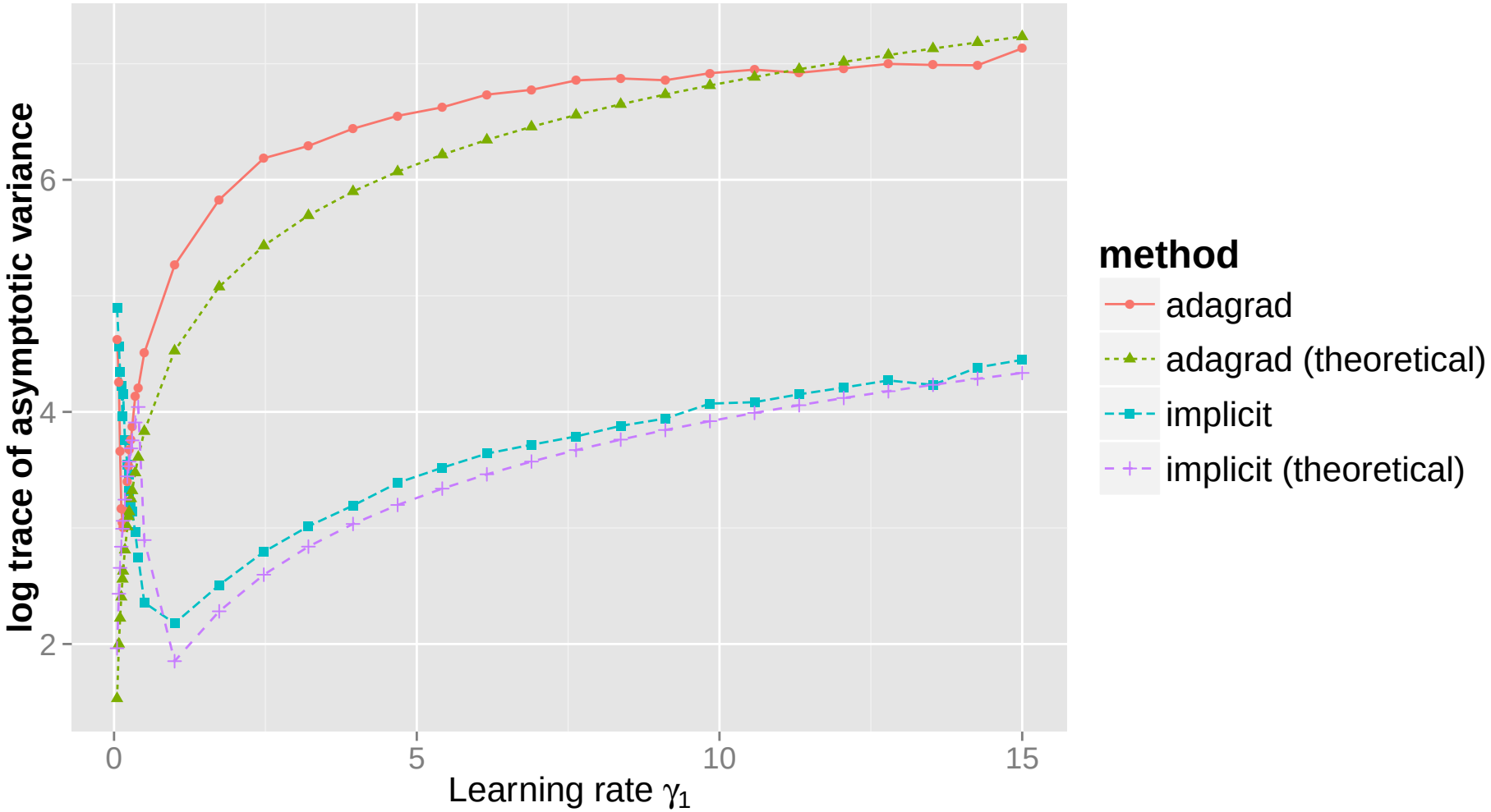
(Toulis & Airoldi, 2015a)

$$\sqrt{\mathbf{n}}\mathbb{V}\text{ar}(\theta_n^{\text{ada}}) \to \frac{\gamma_1}{2}\text{diag}(\mathcal{I}(\theta_\star))^{-1/2}. \tag{4}$$

- AdaGrad is inefficient but (1) holds **regardless** of γ_1.
- In contrast, SGD procedures require $\gamma_1 > 1/(2\lambda_{\text{min}})$ for $O(1/n)$ efficiency.

AdaGrad trade-off: simulation

- $\theta_\star = (2.23, 0.5, 0.1, 0.02, 0.01)^\intercal$; $\lambda_j \in [1, 10]$



Back to main ▷.

Implicit stochastic approximation: implementations

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

1. Run separate RM procedure at each nth 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 θ_n as an estimate of θ^*_{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 $\gamma_n \propto n^{-\gamma}$, $\gamma \in (0,1)$, $\lambda_{\min} > 0$,

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

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

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

$$n \mathbb{V}\text{ar}(\bar{\theta}_n) \to \mathcal{I}(\theta_\star)^{-1}.$$

- $\lambda_{\min} > 0$ 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 ▷.

Bayesian interpretation of implicit methods

- Implicit SGD can be written as

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

- Thus, θ_n^{im} is the *posterior mode* of the Bayesian model,

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

  - Implicit SGD: interpretation of γ_n as information parameter.
  - Explicit SGD: interpretation of γ_n as "step-size".

- First implicit method by Nagumo & Noda (1967); (Slock, 1993)

Back to main ▷.

Connection to proximal methods

- 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.

Incremental proximal gradient

- 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:

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

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

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

- Fairly general way to construct pivotal quantity for $\theta_\star$.
- But we pay price in efficiency.

Back to optimal rates ▷.

The unusual technical challenge of implicit SGD

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

The unusual technical challenge of implicit SGD

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

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

which requires

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

The unusual technical challenge of implicit SGD

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

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

which requires

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

- However, in the implicit procedure

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

we cannot use standard analysis because

$$(Y_n, X_n) \not\perp\!\!\!\perp \theta_n^{\mathsf{im}}.$$

Unusual technical challenge: our approach

- In many statistical models

$$f(Y; X, \theta) \equiv f(Y; X, X^{\intercal}\theta).$$

Unusual technical challenge: our approach

- In many statistical models

$$f(Y; X, \theta) \equiv f(Y; X, X^\intercal \theta).$$

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

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

Unusual technical challenge: our approach

- In many statistical models

$$f(Y; X, \theta) \equiv f(Y; X, X^\intercal \theta).$$

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

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

1. ξ_n is easy to calculate $\Rightarrow$ fast implementation!
2. a.s. bound for ξ_n $\Rightarrow$ avoids conditioning problem since $(Y_n, X_n) \perp\!\!\!\perp \theta_{n-1}^{\mathsf{im}}$.

Proceed with analysis ▷. Back to main ▷.

Almost-sure bound for ξ_n

- Start with

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

Almost-sure bound for ξ_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 θ_{n-1}^{im} yields

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

Almost-sure bound for ξ_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 θ_{n-1}^{im} yields

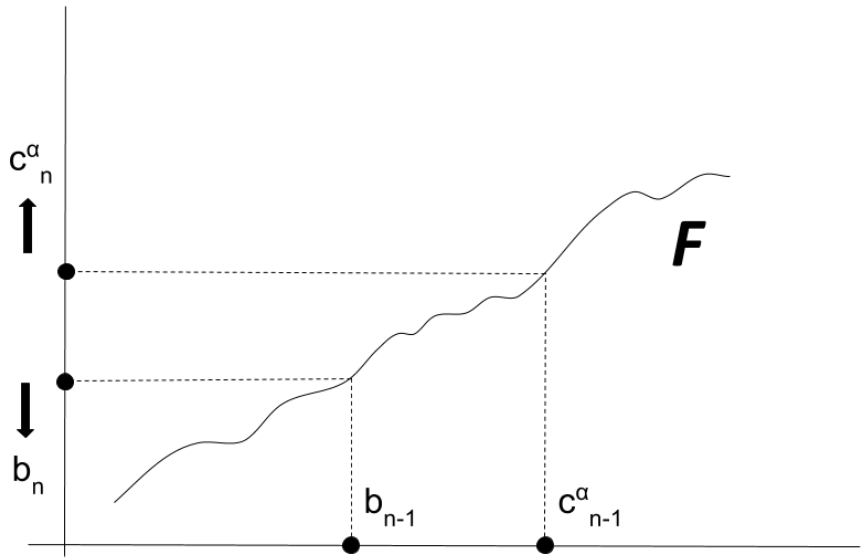
$$\xi_n \geq (1 + \gamma_n s)^{-1} \text{ a.s.}$$

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

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

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

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

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

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 $\alpha_0 n^2 + \alpha_1 n$.

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:

$$\begin{aligned} c_n^\alpha &\geq c_{n-1}^\alpha + n \\ \alpha_0 n^2 + \alpha_1 n &\geq \alpha_0 (n-1)^2 + \alpha_1 (n-1) + n \\ (2\alpha_0 - 1)n + \alpha_1 &\geq \alpha_0 \end{aligned}$$

Thus, $\alpha_0 \geq .5$ and $\alpha_1 \geq \alpha_0$. Therefore,

$$b_n \leq c_n^* = \arg\min_\alpha c_n^\alpha = .5n^2 + .5n = n(n+1)/2$$

Back to main ▷.

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

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 $T(\theta) = \mathbb{E}\left(S|\theta\right)$,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 θ_n converges to point θ_∞ such that

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

Back to main ▷.