Convergence diagnostics for stochastic gradient descent with constant step size

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

Joint work with Jerry Chee (U Chicago, McKinsey)

Econometrics and Statistics
University of Chicago, Booth School of Business

Transient and stationary phase

- Iterative procedures in stochastic optimization are typically comprised of a *transient* phase and a *stationary* phase.
- In the transient phase the procedure converges towards a region of interest.
- During the stationary phase the procedure oscillates in that region, commonly around a single point.
- Understanding when the phase transition happens is crucial for implementation and for improving empirical performance.
- Our focus here will be stochastic gradient descent (SGD) procedures, but our results may be more general.

Stochastic gradient descent (SGD)

- Statistical estimation gave a new form of optimization problems:

$$\theta_\star = \arg\min_\theta \ell(\theta) = \arg\min_\theta \sum_{i=1}^N l_i(\theta),$$

where ℓ is loss function; l_i is loss for ith datapoint only.

Stochastic gradient descent (SGD)

- Statistical estimation gave a new form of optimization problems:

$$\theta_\star = \arg\min_\theta \ell(\theta) = \arg\min_\theta \sum_{i=1}^{N} l_i(\theta),$$

where ℓ is loss function; l_i is loss for ith datapoint only.

- Classical optimization methods, such as gradient descent,

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

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

Stochastic gradient descent (SGD)

- Statistical estimation gave a new form of optimization problems:

$$\theta_\star = \arg\min_\theta \ell(\theta) = \arg\min_\theta \sum_{i=1}^N l_i(\theta),$$

where ℓ is loss function; l_i is loss for ith datapoint only.

- Classical optimization methods, such as gradient descent,

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

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

- SGD has emerged as one of the most versatile optimization methods:

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

where $J \sim \text{Unif}[1, 2, \ldots, N]$.

- By SA theory ▷, $\theta_n \to \theta_\infty$ such that: $\mathbb{E}\left(\nabla l_J(\theta_\infty)\right) = 0 \Rightarrow \theta_\infty = \theta_\star$.

SGD with constant step size

- While decreasing step size converges (in theory) it produces several problems: (1) sensitivity to misspecification; (2) slow rate of convergence – $O(1/N)$.
- SGD with **constant** step size behaves differently:

$$\theta_n = \theta_{n-1} - \gamma \nabla l_J(\theta_{n-1}).$$

  - "Convergence" is much faster...
  - ..but not real convergence! Actually converges to region of radius $O(\sqrt{\gamma})$ that contains $\theta_\star$, and then oscillates in this region.
- We will try to identify when SGD reaches the convergence region:

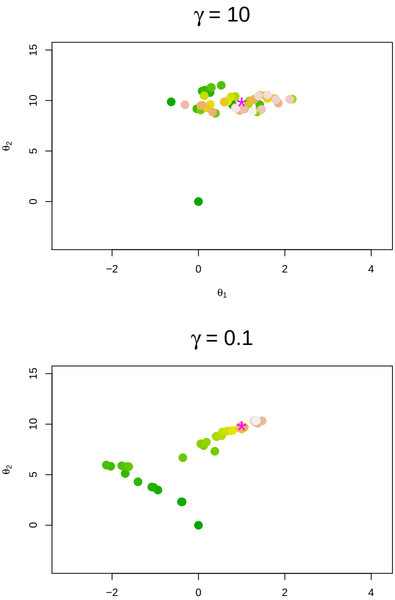
SGD with constant step size

- While decreasing step size converges (in theory) it produces several problems: (1) sensitivity to misspecification; (2) slow rate of convergence – $O(1/N)$.
- SGD with **constant** step size behaves differently:

$$\theta_n = \theta_{n-1} - \gamma \nabla l_J(\theta_{n-1}).$$

  - "Convergence" is much faster...
  - ..but not real convergence! Actually converges to region of radius $O(\sqrt{\gamma})$ that contains $\theta_\star$, and then oscillates in this region.
- We will try to identify when SGD reaches the convergence region:
  - Pointless to run the procedure beyond that point.
  - Can improve the procedure by detecting convergence and then updating it (e.g., decrease the step size).

Illustration: SGD with constant step size



An intuition for such behavior is in the following meta-theorem:

Theorem (Zhang, 2004); (Moulines and Bach, 2011); (Needell et. al., 2014)

There are positive constants A_γ, B such that, for every n, it holds that

$$\mathbb{E}\left(||\theta_n - \theta_\star||^2\right) \leq \mathbb{E}\left(||\theta_0 - \theta_\star||^2\right) e^{-A_\gamma n} + B\gamma.$$

- For example, $A_\gamma \approx \gamma\mu/4 - \gamma^2 L^2$, where μ, L are strong convexity and Lipschitz constant of expected loss, resp; $B = \sigma^2/\mu$, where σ^2 is noise level.

An intuition for such behavior is in the following meta-theorem:

Theorem (Zhang, 2004); (Moulines and Bach, 2011); (Needell et. al., 2014)

There are positive constants A_γ, B such that, for every n, it holds that

$$\mathbb{E}\left(||\theta_n - \theta_\star||^2\right) \leq \mathbb{E}\left(||\theta_0 - \theta_\star||^2\right) e^{-A_\gamma n} + B\gamma.$$

- For example, $A_\gamma \approx \gamma\mu/4 - \gamma^2 L^2$, where μ, L are strong convexity and Lipschitz constant of expected loss, resp; $B = \sigma^2/\mu$, where σ^2 is noise level.
- **Transient phase**: SGD forgets initial conditions exponentially fast.
- **Stationary phase:** SGD oscillates around $\theta_\star$ at a region of radius $O(\sqrt{\gamma})$.
- Trade-off: large γ speeds up convergence but increases oscillation radius; small γ decreases the radius but convergence is slower.

An intuition for such behavior is in the following meta-theorem:

Theorem (Zhang, 2004); (Moulines and Bach, 2011); (Needell et. al., 2014)

There are positive constants A_γ, B such that, for every n, it holds that

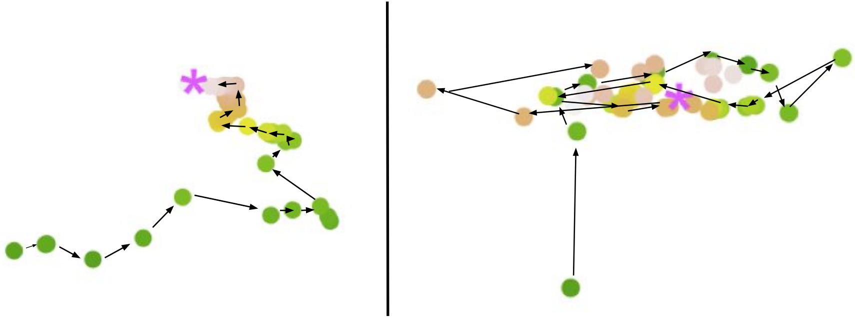
$$\mathbb{E}\left(||\theta_n - \theta_\star||^2\right) \le \mathbb{E}\left(||\theta_0 - \theta_\star||^2\right) e^{-A_\gamma n} + B\gamma.$$

- For example, $A_\gamma \approx \gamma\mu/4 - \gamma^2 L^2$, where μ, L are strong convexity and Lipschitz constant of expected loss, resp; $B = \sigma^2/\mu$, where σ^2 is noise level.
- **Transient phase**: SGD forgets initial conditions exponentially fast.
- **Stationary phase:** SGD oscillates around $\theta_\star$ at a region of radius $O(\sqrt{\gamma})$.
- Trade-off: large γ speeds up convergence but increases oscillation radius; small γ decreases the radius but convergence is slower.
- Despite valuable theoretical insights such results offer limited guidance in practice for detecting convergence (bound may not be tight; parameters μ, L, σ^2 hard to estimate).

Related work

- The idea of transient/stationary phases (also known as search/convergence phases) has been expressed before (e.g., Murata, 1998).
- In optimization, a typical approach is to stop when $||\theta_n - \theta_{n-1}||$ is small according to some threshold, or when updates of the loss function have reached machine precision (Ermoliev and Wets, 1998; Bottou et. al., 2016). Ignores noise from stochastic gradients.
- Large literature on convergence diagnostics of Monte Carlo Markov Chains (Cowles, 1996). Different setting but shares common characteristics with our problem here.
- Pflug has made seminal contributions in the theory of stopping times in stochastic approximations (1998, 1990). Our work here is **heavily** influenced by Pflug's work.

Pflug’s convergence diagnostic (high-level idea)



- (Left) In transient phase gradient is auto-correlated: successive gradients generally point to same direction.
- (Right) In convergence phase successive gradients are more likely to point to opposite direction.
- Running average of inner product of successive gradients will thus be our test statistic.

Convergence diagnostic algorithm

Let ∇l_n = stoch. gradient at nth iteration (depends on sampled y_n and θ_{n-1}).

$S_0 \leftarrow 0$
$\theta_1 \leftarrow \theta_0 - \gamma \nabla l_1$
for all $n \in \{2, 3, \cdots\}$ **do**
    $\theta_n \leftarrow \theta_{n-1} - \gamma \nabla l_n.$
    $S_n \leftarrow S_{n-1} + \nabla l_n^\top \nabla l_{n-1}$ *#running sum of inner product*
    if $n >$ `burnin` and $S_n < 0$ **then**
        return n *#declare convergence*
    end if
end for

- Variable `burnin` = $O(1/\gamma)$.
- Several variations of the algorithm are possible; e.g., discount old iterations in running sum.

Quadratic loss model: first intuition

Let x =features, y = outcomes; we focus on quadratic loss where

$$\ell(y, x; \theta) = (1/2)(y - x^\top \theta)^2 \text{ and } \nabla \ell(y, x; \theta) = -(y - x^\top \theta)x.$$

- Suppose that $\theta_0 = \theta_\star$. Let $y_n - x_n^\top \theta_\star = \varepsilon_n$, where ε_n are zero-mean r.v. given x_n. Then,

  $$\theta_1 = \theta_\star + \gamma(y_1 - x_1^\top \theta_\star)x_1 = \theta_\star + \gamma \varepsilon_1 x_1,$$

  from which it follows that

  $$S_2 - S_1 = (y_2 - x_2^\top \theta_1)(y_1 - x_1^\top \theta_0)x_2^\top x_1 = (\varepsilon_2 - \gamma \varepsilon_1 x_2^\top x_1)\varepsilon_1 x_2^\top x_1.$$

  $$\mathbb{E}\left(S_2 - S_1\right) = -\gamma \mathbb{E}\left(\varepsilon_1^2\right) \mathbb{E}\left((x_2^\top x_1)^2\right) < 0. \quad (1)$$

Quadratic loss model: first intuition

Let x =features, y = outcomes; we focus on quadratic loss where

$$\ell(y, x; \theta) = (1/2)(y - x^\top \theta)^2 \text{ and } \nabla \ell(y, x; \theta) = -(y - x^\top \theta)x.$$

- Suppose that $\theta_0 = \theta_\star$. Let $y_n - x_n^\top \theta_\star = \varepsilon_n$, where ε_n are zero-mean r.v. given x_n. Then,

$$\theta_1 = \theta_\star + \gamma(y_1 - x_1^\top \theta_\star)x_1 = \theta_\star + \gamma \varepsilon_1 x_1,$$

  from which it follows that

$$S_2 - S_1 = (y_2 - x_2^\top \theta_1)(y_1 - x_1^\top \theta_0)x_2^\top x_1 = (\varepsilon_2 - \gamma \varepsilon_1 x_2^\top x_1)\varepsilon_1 x_2^\top x_1.$$

$$\mathbb{E}\left(S_2 - S_1\right) = -\gamma \mathbb{E}\left(\varepsilon_1^2\right) \mathbb{E}\left((x_2^\top x_1)^2\right) < 0. \quad (1)$$

- Thus, the diagnostic is decreased in expectation, and by LLN (and a property of upper-boundedness) it will *eventually* become negative.

Theorem

For quadratic loss, let x_1 *and* x_2 *be two iid vectors from the distribution of* x*, and define:* $\sigma^2 = \mathbb{E}\left((y - x^\top \theta_\star)^2\right)$*;* $c^2 = \mathbb{E}\left((x_1^\top x_2)^2\right)$*;* $C = \mathbb{E}\left(x_1 x_2^\top (x_1^\top x_2)\right)$*;* $D = \mathbb{E}\left(x_1 x_1^\top (x_1^\top x_2)^2\right)$*, and suppose that all are finite. Then, for* $\gamma > 0$*,*

$$\begin{aligned}\Delta_n(\theta) &= \mathbb{E}\left(S_{n+2} - S_{n+1} | \theta_n = \theta\right) \\ &= (\theta - \theta_\star)^\top (C - \gamma D)(\theta - \theta_\star) - \gamma c^2 \sigma^2.\end{aligned}$$

Theorem

For quadratic loss, let x_1 *and* x_2 *be two iid vectors from the distribution of* x*, and define:* $\sigma^2 = \mathbb{E}\left((y - x^\top\theta_\star)^2\right)$*;* $c^2 = \mathbb{E}\left((x_1^\top x_2)^2\right)$*;* $C = \mathbb{E}\left(x_1 x_2^\top (x_1^\top x_2)\right)$*;* $D = \mathbb{E}\left(x_1 x_1^\top (x_1^\top x_2)^2\right)$*, and suppose that all are finite. Then, for* $\gamma > 0$*,*

$$\begin{aligned}\Delta_n(\theta) &= \mathbb{E}\left(S_{n+2} - S_{n+1} | \theta_n = \theta\right)\\ &= (\theta - \theta_\star)^\top (C - \gamma D)(\theta - \theta_\star) - \gamma c^2 \sigma^2.\end{aligned}$$

- Boundary surface of expected sign increase of test statistic is an ellipse.
- When $\theta = \theta_\star$ we have expected decrease (good!).
- Interesting dynamics:
  - Bias term contributes positive values to test statistic. Reasonable because bias pushes SGD iterates to $\theta_\star$.
  - Error term contributes negative values to statistic. Reasonable because error pushes SGD iterates away from $\theta_\star$.

Assume $y \sim \mathcal{N}(x^\top \theta_\star, \sigma^2)$; $x \sim \mathcal{N}(0, I_2)$; $\theta_\star = (0.47, 0.22)$; $\sigma^2 = 3$.

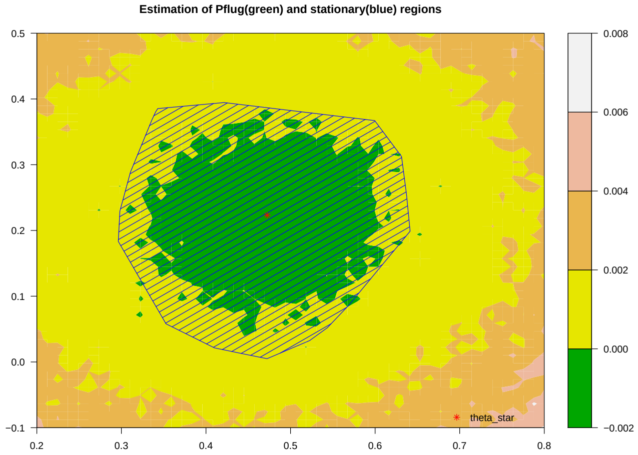
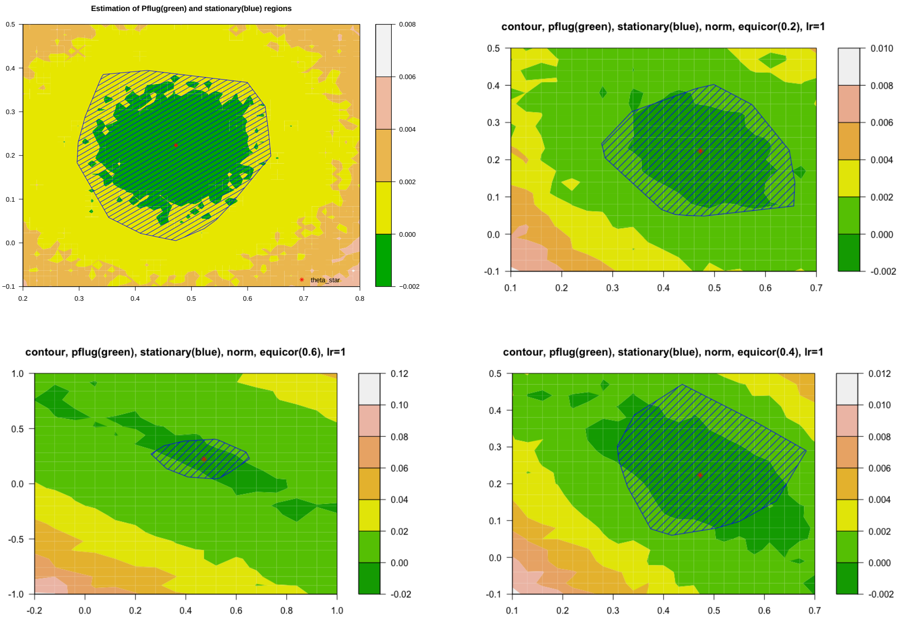
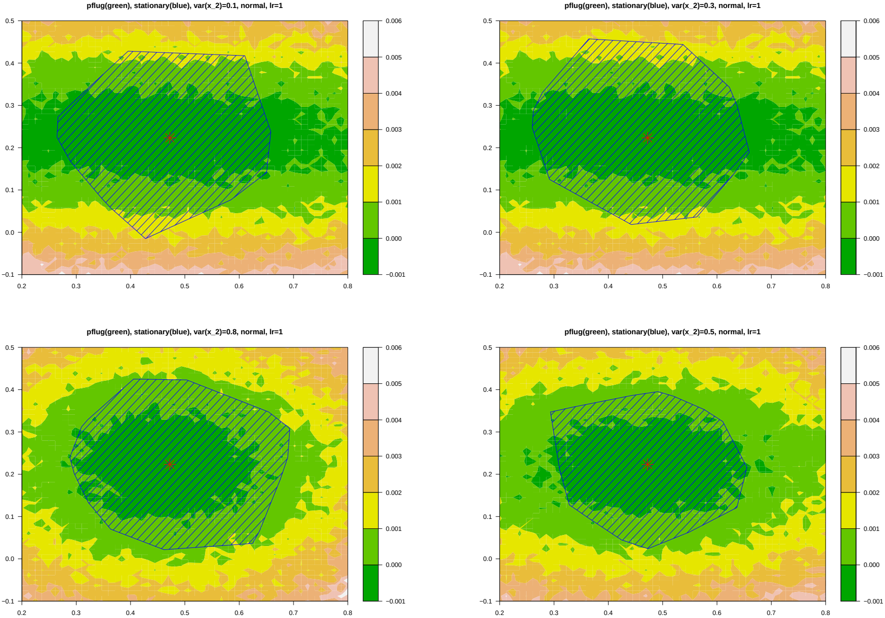


Figure: Green center: Pflug diagnostic decreased in expectation. Blue polygon: oscillation region of SGD iterates (empirically calculated). Color legend: values of expected increase (or decrease) of the diagnostic.

Equicorrelated case: $\mathrm{cor}(x_1, x_2) = \rho \in [0, 0.2, 0.4, 0.6]$.



Ill conditioning: $\text{Var}(x_1) = 1, \text{Var}(x_2) \in [0.1, 0.3, 0.5, 0.8]$.



Simulated study

- $p = 20$ dimensions; $\theta_{\star,j} = 10e^{-0.7j}$; $\sigma = 3$; $N = 5000$ data points;
- Let $E_n = ||\theta_n - \theta_\star||^2$ and τ be when the test diagnostic is activated.
- We store $(\gamma, \tau, E_0, E_{\tau/2}, E_{2\tau})$.

Simulated study

- $p = 20$ dimensions; $\theta_{\star,j} = 10e^{-0.7j}$; $\sigma = 3$; $N = 5000$ data points;
- Let $E_n = ||\theta_n - \theta_\star||^2$ and τ be when the test diagnostic is activated.
- We store $(\gamma, \tau, E_0, E_{\tau/2}, E_{2\tau})$.

	$E_{\tau/2} = \beta_{\tau/2} E_0 + \varepsilon$	$E_{2\tau} = \beta_{2\tau} E_0 + \varepsilon$
γ	$\beta_{\tau/2}$	$\beta_{2\tau}$
0.02	0.17 **	0.01 .
0.05	0.20 ***	-0.008
0.1	0.09 **	-0.0007
0.2	0.06 **	0.005
0.5	0.09 ***	-0.008
1.0	0.06 *	0.02 *
2.0	0.06 **	0.009
5.0	0.07 **	-0.012

Simulated study

- $p = 20$ dimensions; $\theta_{\star,j} = 10e^{-0.7j}$; $\sigma = 3$; $N = 5000$ data points;
- Let $E_n = ||\theta_n - \theta_\star||^2$ and τ be when the test diagnostic is activated.
- We store $(\gamma, \tau, E_0, E_{\tau/2}, E_{2\tau})$.

	$E_{\tau/2} = \beta_{\tau/2} E_0 + \varepsilon$	$E_{2\tau} = \beta_{2\tau} E_0 + \varepsilon$
γ	$\beta_{\tau/2}$	$\beta_{2\tau}$
0.02	0.17 **	0.01 .
0.05	0.20 ***	−0.008
0.1	0.09 **	−0.0007
0.2	0.06 **	0.005
0.5	0.09 ***	−0.008
1.0	0.06 *	0.02 *
2.0	0.06 **	0.009
5.0	0.07 **	−0.012

- Table shows that diagnostic behaves as intended: conditional on activated diagnostic the distance to $\theta_\star$ is uncorrelated with initial distance.

Sensitivity and Implicit update

- Pflug diagnostic and main SGD procedure are sensitive to misspecification of step size γ.
- One way to alleviate such sensitivities is to use the SGD procedure with an **implicit update** (ISGD):

$$\boldsymbol{\theta_n} = \theta_{n-1} - \gamma \nabla l_J(\boldsymbol{\theta_n}).$$

- For quadratic loss the implicit update is equivalent to:

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

Sensitivity and Implicit update

- Pflug diagnostic and main SGD procedure are sensitive to misspecification of step size γ.
- One way to alleviate such sensitivities is to use the SGD procedure with an **implicit update** (ISGD):

$$\boldsymbol{\theta_n} = \theta_{n-1} - \gamma \nabla l_J(\boldsymbol{\theta_n}).$$

- For quadratic loss the implicit update is equivalent to:

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

- **Note:** Implicit update is more easily applicable than usually assumed in practice (Toulis et.al., 2014) – straightforward, and essentially costless computationally, for generalized linear models, M-estimation, hazards.
- Implicit SGD procedures are statistically equivalent to explicit ones, but remarkably more robust numerically (Toulis and Airoldi, 2017).

(more details on implicit computation ▷; relation to Bayesian interpretation ▷, or proximal optimization ▷)

Implicit update

Theorem

Let $\lambda_\gamma = \mathbb{E}\left(1/(1+\gamma||x||^2)\right) \in (0,1]$ *and* $\Delta_n^{\text{im}}(\theta) = \mathbb{E}\left(S_{n+2} - S_{n+1} | \theta_n = \theta\right)$. *Then, it holds that*

$$\Delta_n^{\text{im}}(\theta) = a_\gamma \Delta_n(\theta) + b_\gamma \left[(\theta - \theta_\star)^\top D(\theta - \theta_\star) + \sigma^2 c^2\right],$$

where $\Delta_n(\theta)$ *is the expected increase for the explicit update,* $a_\gamma = \lambda_\gamma^2$*, and* $b_\gamma = \gamma\lambda_\gamma^2(1-\lambda_\gamma)$.

Implicit update

Theorem

Let $\lambda_\gamma = \mathbb{E}\left(1/(1+\gamma||x||^2)\right) \in (0,1]$ *and* $\Delta_n^{\text{im}}(\theta) = \mathbb{E}\left(S_{n+2} - S_{n+1} | \theta_n = \theta\right)$. *Then, it holds that*

$$\Delta_n^{\text{im}}(\theta) = a_\gamma \Delta_n(\theta) + b_\gamma \left[(\theta - \theta_\star)^\top D(\theta - \theta_\star) + \sigma^2 c^2\right],$$

where $\Delta_n(\theta)$ *is the expected increase for the explicit update,* $a_\gamma = \lambda_\gamma^2$, *and* $b_\gamma = \gamma\lambda_\gamma^2(1-\lambda_\gamma)$.

- Consider, for example, $\gamma = \infty$ so that $\lambda_\gamma = 0$. In classical SGD the diagnostic increases without bound and convergence fails.
- With implicit update we have $\Delta_n^{\text{im}}(\theta) \approx 0$, and convergence may happen.
- In contrast, when $\gamma \approx 0$ then $\lambda_\gamma \approx 1$ and so $\Delta_n^{\text{im}}(\theta) \approx \Delta_n(\theta)$, and implicit diagnostic behaves as the explicit one.

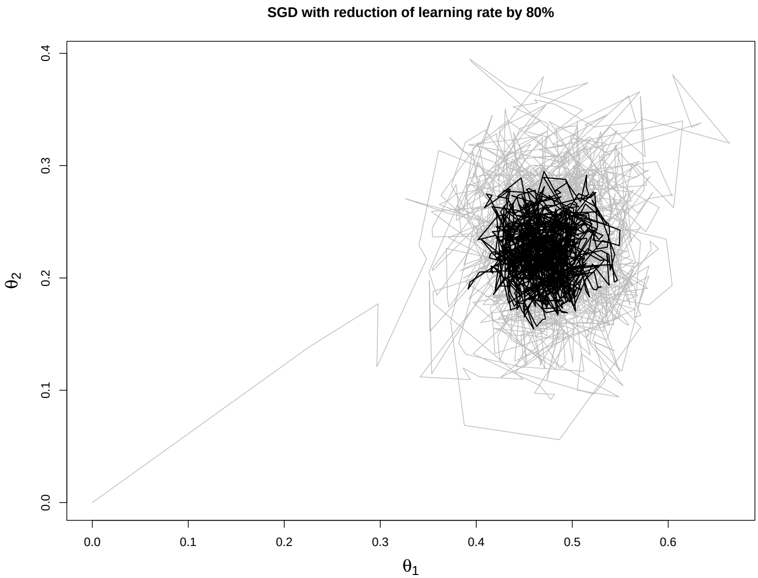
Summary regarding Pflug diagnostic

- Diagnostic activation region can be obtained in closed form for quadratic loss (and, more generally, for generalized linear models – not shown here).
- Activation region *empirically* coincides with actual stationary region.
- Distance $||\theta_n - \theta_\star||^2$ *empirically* uncorrelated with initial distance conditional on diagnostic being activated.
- Implicit update offers a more reliable version of the diagnostic.
- Performance is hurt by ill conditioning and poor initialization (ongoing work).

Application: ISGD$^{1/2}$

- We discuss one application of the diagnostic to define a version of SGD that converges to $\theta_\star$ in linear time.
- The idea is simply to reduce the step size (e.g., halve, $\gamma \leftarrow \gamma/2$) each time convergence is detected.
- We call this procedure ISGD$^{1/2}$. We do not have a theoretical analysis of its performance, only of its parts (ISGD with constant step size and Pflug diagnostic).

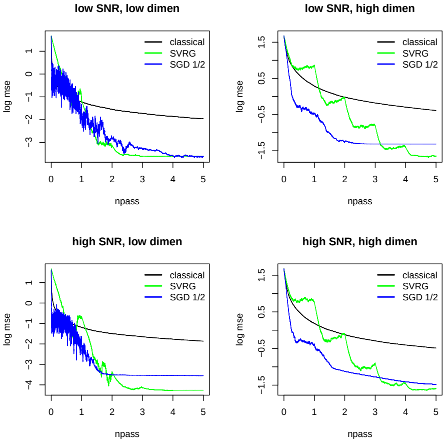
Illustration of ISGD$^{1/2}$



Experimental setup for ISGD$^{1/2}$

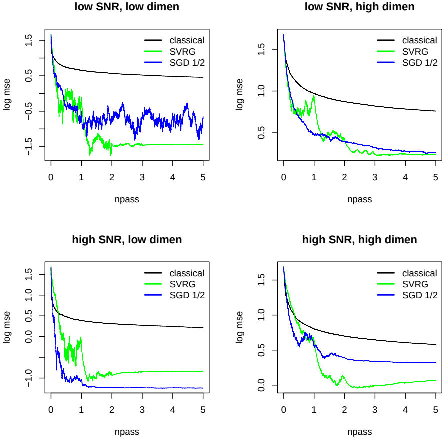
- We compare ISGD$^{1/2}$ against SVRG and ISGD on simulated data.
- We consider high and low dimension settings as $p = 150$ and $p = 10$, respectively.
- We consider high and low signal to noise ratio (SNR) settings as SNR $= 5$ and SNR $= 2$, where SNR $=$ trace$(\mathbb{V}\text{ar}(x))/p\mathbb{V}\text{ar}(y|x)$.
- We fix $\theta_\star$ such that $\theta_{\star,j} = 10e^{-0.75j}$; we set $N = 5000$.
- We sample $x_i \sim \mathcal{N}_p(0, I)$, where $i = 1, 2, \ldots N$.
- We sample $y_i \sim \mathcal{N}(x_i^\top \theta_\star, \sigma^2)$ for normal model, and $y_i \sim$ Binom$(\exp(x_i^\top \theta_\star)/(1 + \exp(x_i^\top \theta_\star))$ for logistic model.

ISGD$^{1/2}$ on normal model



- ISGD$^{1/2}$ attains comparable performance to SVRG. Still, SVRG is better here, overall.
- We believe we can improve ISGD$^{1/2}$ if we reduce Type-I error rate of the diagnostic.
- Type-I errors lead to very small step sizes early in the procedure, which slows us down.
- Halving the learning may also be too aggressive.

ISGD$^{1/2}$ on logistic regression model



- Mixed picture again. ISGD$^{1/2}$ still comparable to SVRG.
- In high SNR-few dimensions, ISGD$^{1/2}$ achieves consistently better performance than SVRG.
- Larger `burnin` period or discounting the step size less aggressively can also help here.
- We plan on addressing such tuning issues in future work, both theoretically and empirically.

Concluding remarks

- Convergence diagnostics are useful for stopping SGD when necessary, and building improved variants.
- Type-I and Type-II error properties of Pflug diagnostic still unknown (and challenging to analyze!).
- Future work can focus also on analysis conditional on diagnostic being activated (Table results in this talk).
- Also focus more on $\text{ISGD}^{1/2}$ that worked very well in experiments. Convergence rate analysis? Tuning?
- Parallelization is unexplored so far. One idea is to run parallel $\text{ISGD}^{1/2}$ chains and aggregate iterates. At stationarity we expect iterates from different chains to be uncorrelated with each other.

THANK YOU!

- Chee J, Toulis P, “Convergence diagnostics for stochastic gradient descent with constant step size” (2017, arxiv)
- Toulis P, Airoldi EM, “Asymptotic and finite-sample properties of estimators based on stochastic gradients.” (2017, Annals of Statistics)
- Toulis P, Rennie J, Airoldi EM, “Statistical analysis of stochastic gradient methods for generalized linear models”, (2014, Int’l Conference in Machine Learning, ICML)

Other references

- Bottou, Leon, Frank E. Curtis, and Jorge Nocedal. “Optimization methods for large-scale machine learning.” arXiv preprint arXiv:1606.04838 (2016).
- Cowles, Mary Kathryn, and Bradley P. Carlin. “Markov chain Monte Carlo convergence diagnostics: a comparative review.” Journal of the American Statistical Association 91.434 (1996): 883-904.
- Ermoliev, Yu M., and RJ-B. Wets. “Numerical techniques for stochastic optimization.” Springer-Verlag, 1988.
- Moulines, Eric, and Francis R. Bach. “Non-asymptotic analysis of stochastic approximation algorithms for machine learning.” Advances in Neural Information Processing Systems. 2011.
- Murata, Noboru. “A statistical study of on-line learning.” Online Learning and Neural Networks. Cambridge University Press, Cambridge, UK (1998): 63-92. APA
- Needell, Deanna, Rachel Ward, and Nati Srebro. “Stochastic gradient descent, weighted sampling, and the randomized kaczmarz algorithm.” Advances in Neural Information Processing Systems. 2014.
- Pflug, G. Ch. “Stepsize rules, stopping times and their implementation in stochastic quasigradient algorithms.” numerical techniques for stochastic optimization (1988): 353-372.
- Pflug, Georg Ch. “Non-asymptotic confidence bounds for stochastic approximation algorithms with constant step size.” Monatshefte fur Mathematik 110.3 (1990): 297-314.
- Zhang, Tong. “Solving large scale linear prediction problems using stochastic gradient descent algorithms.” Proceedings of the twenty-first international conference on Machine learning. ACM, 2004.

Backup slides

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

Backup slides

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

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

Backup slides

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

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

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

Backup slides

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

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

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

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

Backup slides

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

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

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

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

$$\ldots$$

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

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

Efficient computation of implicit updates

Suppose that $\nabla l(\theta) = s(y, x^\top \theta)x$, and ignore step size γ_n. Then,

Efficient computation of implicit updates

Suppose that $\nabla l(\theta) = s(y, x^\top \theta)x$, and ignore step size γ_n. Then,

$$\theta_n^{\text{im}} = \theta_{n-1}^{\text{im}} + s(y_n, x_n^\top \theta_n^{\text{im}})x_n \tag{2}$$

$$= \theta_{n-1}^{\text{im}} + \xi s(y_n, x_n^\top \theta_{n-1}^{\text{im}})x_n \tag{3}$$

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

Efficient computation of implicit updates

Suppose that $\nabla l(\theta) = s(y, x^\top \theta)x$, and ignore step size γ_n. Then,

$$\theta_n^{\text{im}} = \theta_{n-1}^{\text{im}} + s(y_n, x_n^\top \theta_n^{\text{im}})x_n \tag{2}$$
$$= \theta_{n-1}^{\text{im}} + \xi s(y_n, x_n^\top \theta_{n-1}^{\text{im}})x_n \tag{3}$$
$$\triangleq \theta_{n-1}^{\text{im}} + a_n x_n. \tag{4}$$

Equate the two scales:

$$a_n = s(y_n, x_n^\top \theta_n^{\text{im}}) \qquad \text{[by setting (1) = (3)]}$$
$$= s(y_n, x_n^\intercal \theta_{n-1}^{\text{im}} + ||x_n||^2 a_n). \qquad \text{[by substituting } \theta_n^{\text{im}} \text{ with (3)]}$$

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

$$u = s(y, a + cu)$$

where $c > 0$. It follows that $u \in [\min(0, s(y, a)), \max(0, s(y, a))]$.

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

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

Go back ▷.

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

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

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

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

Optimal efficiency: second-order SGD

Theorem (Toulis & Airoldi, 2015a)

Consider the second-order implicit SGD procedure

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

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

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

- 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

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

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

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

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

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

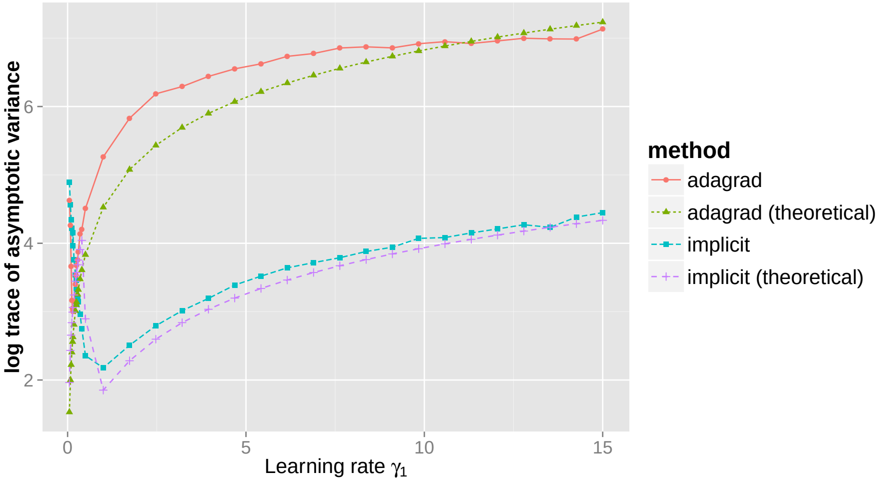
(Toulis & Airoldi, 2015a)

$$\sqrt{\mathbf{n}} \mathbb{V}\text{ar}(\theta_n^{\text{ada}}) \to \frac{\gamma_1}{2} \text{diag}(\mathcal{I}(\theta_\star))^{-1/2}. \tag{5}$$

- AdaGrad is inefficient but (1) holds **regardless** of γ_1.
- In contrast, SGD procedures require $\gamma_1 > 1/(2\mu)$ 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, $k = 1, 2, \ldots$

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

   - $x_k \rightarrow \theta^*_{n-1}$ (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 $\gamma_n \propto n^{-\gamma}$, $\gamma \in (0, 1)$, $\mu > 0$,

$$\theta_n^{\text{im}} = \theta_{n-1}^{\text{im}} + \gamma_n \nabla \log f(Y_n; X_n, \theta_n^{\text{im}})$$
$$\bar{\theta}_n = \frac{1}{n} \sum_{i=1}^{n} \theta_i^{\text{im}}.$$

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

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

- $\mu > 0$ 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)

Go back ▷.

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 back ▷.

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:

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

- If $\gamma_1 >> 1/(2\mu)$,

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

- 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 $\mathbb{E}\left(||\theta_n^{\text{ex}} - \theta_\star||^2\right)$.
- A crucial property is the concavity of

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

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 $\mathbb{E}\left(||\theta_n^{\text{ex}} - \theta_\star||^2\right)$.
- 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

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

we cannot use standard analysis because

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

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,

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

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,

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

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

Proceed with analysis ▷. Back to main ▷.

Almost-sure bound for ξ_n

- Start with

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

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

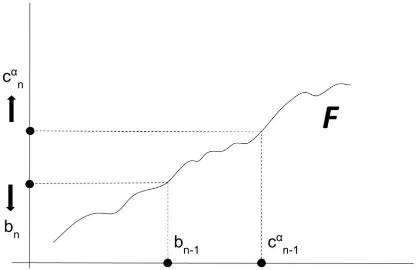
$$\xi_n \geq (1 + \gamma_n s)^{-1} \text{ a.s.}$$

- 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 ▷.