AIE6003: Optimization for Machine Learning

Xiaopeng Li
CUHK(SZ)
SAI

March 4, 2026

Outline

Stochastic Optimization

Convergence of SGD

Variance Reduction

Introduction

Population model:

$$\min_{x \in \mathbb{R}^d} \; f(x) := \mathbb{E}_{\xi \sim \mathcal{D}}[F(x, \xi)].$$

Empirical risk minimization (ERM): with samples $\{\xi_i\}_{i=1}^N$,

$$f(x) = \frac{1}{N} \sum_{i=1}^{N} f_i(x), \quad f_i(x) := F(x, \xi_i).$$

Motivation: evaluating $\nabla f(x)$ is expensive when N is huge.

Idea: use cheap unbiased gradients from a single (or mini-batch) sample.

Stochastic Gradient Descent (SGD)

Population form: we query x and draw an independent sample $\xi \sim \mathcal{D}$, returning $g(x, \xi) \in \partial F(x, \xi)$ such that

$$\mathbb{E}\left[g(x, \xi) \mid x\right] = \nabla f(x) \quad \text{(unbiased under i.i.d. sampling)}.$$

Finite-sum form: $f(x) = \frac{1}{N} \sum_{i=1}^{N} f_i(x)$ with $g_t = \nabla f_{i_t}(x_t)$.

- if $i_t \sim \mathrm{Unif}\{1, \ldots, N\}$ with replacement, then

$$\mathbb{E}[g_t \mid x_t] = \nabla f(x_t), \quad x_{t+1} = x_t - \eta_t g_t.$$

Other sampling schemes:

- **without replacement (random reshuffling):** indices are dependent within an epoch; typically biased, i.e., $\mathbb{E}[g_t \mid x_t] \neq \nabla f(x_t)$.
- **cyclic:** deterministic incremental gradient.

A Sanity Check: Averaging via SGD

Toy problem: given points $p_1, \ldots, p_N \in \mathbb{R}^d$, minimize

$$f(x) = \frac{1}{2N} \sum_{i=1}^{N} \|x - p_i\|_2^2 \implies \nabla f_i(x) = x - p_i.$$

Closed-form optimum:

$$x^* = \arg\min_x f(x) = \frac{1}{N} \sum_{i=1}^{N} p_i \quad \text{(the mean)}.$$

Cyclic SGD with $\eta_t = \frac{1}{t+1}$ and $x_0 = 0$:

$$x_{t+1} = x_t - \frac{1}{t+1}(x_t - p_{t+1}).$$

Claim: $x_t = \frac{1}{t} \sum_{i=1}^{t} p_i$ (running average).

Proof Sketch

Induction:

- Base: $x_1 = x_0 - 1 \cdot (0 - p_1) = p_1$.
- Step: assume $x_t = \frac{1}{t} \sum_{i=1}^{t} p_i$.

Compute the update:

$$x_{t+1} = x_t - \tfrac{1}{t+1}(x_t - p_{t+1}) = \tfrac{t}{t+1} x_t + \tfrac{1}{t+1} p_{t+1}$$
$$= \tfrac{t}{t+1} \left(\tfrac{1}{t} \sum_{i=1}^{t} p_i \right) + \tfrac{1}{t+1} p_{t+1} = \tfrac{1}{t+1} \sum_{i=1}^{t+1} p_i.$$

Takeaway: step sizes can "encode" useful averaging.

Empirical Risk Minimization (ERM)

Data: $(x_i, y_i) \sim \mathcal{D}$ i.i.d., features $x_i \in \mathcal{X}$ and labels $y_i \in \mathcal{Y}$.

Goal: learn a predictor $h : \mathcal{X} \to \mathcal{Y}$ by minimizing risk

$$R[h] := \mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell(h(x), y)] \approx R_{\text{train}}[h] := \frac{1}{N}\sum_{i=1}^{N} \ell(h(x_i), y_i).$$

Optimization view:

$$\min_{\theta} \ \frac{1}{N}\sum_{i=1}^{N} \ell(h_\theta(x_i), y_i) + \lambda \cdot \Omega(\theta).$$

How SGD enters ERM:

$$\mathcal{D} \xrightarrow{\text{i.i.d.}} \{(x_i, y_i)\}_{i=1}^{N}, \quad f(\theta) = \frac{1}{N}\sum_{i=1}^{N} \ell(h_\theta(x_i), y_i) + \lambda\,\Omega(\theta),$$

$$\theta_{t+1} = \theta_t - \eta_t\, \nabla_\theta \ell(h_{\theta_t}(x_{i_t}), y_{i_t}) \quad \text{with } i_t \sim \text{Unif}\{1, \ldots, N\}.$$

Perceptron as Stochastic Update

Binary classification: $y_i \in \{-1, +1\}$, linear score $h_w(x) = \langle w, x \rangle$.

Margin violation: an example is "bad" if $y_i \langle w, x_i \rangle < 1$.

Stochastic update: pick $i_t \sim \mathrm{Unif}\{1, \ldots, N\}$ and set

$$w_{t+1} = (1 - \lambda) w_t + \eta_t \, y_{i_t} x_{i_t} \cdot \mathbf{1}\{y_{i_t} \langle w_t, x_{i_t} \rangle < 1\} \, .$$

Interpretation:

- update only when the margin is violated;
- shrinkage $(1 - \lambda) w_t$ mimics ℓ_2 regularization;
- can be viewed as SGD on hinge-loss + regularization.

Support Vector Machine

Regularized hinge-loss objective:

$$\min_{w \in \mathbb{R}^d} \ \frac{1}{N} \sum_{i=1}^{N} \max(1 - y_i \langle w, x_i \rangle, 0) + \lambda \|w\|_2^2.$$

Why hinge loss?

- convex surrogate for the 0-1 loss;
- enforces a unit-margin when separable;
- subgradients are cheap: if $y_i \langle w, x_i \rangle < 1$, subgradient is $-y_i x_i$.

SGD step: sample i_t and take a (sub)gradient step on the sampled hinge-loss term.

Outline

Stochastic Optimization

Convergence of SGD

Variance Reduction

Analysis Template

Setting: minimize $f(x) = \mathbb{E}[F(x, \xi)]$, access only to samples $g(x, \xi)$.

Classical assumptions:

- Unbiasedness: $\mathbb{E}[g(x, \xi)] = \nabla f(x)$.
- Bounded second moment:
$$\sup_x \mathbb{E} \|g(x, \xi)\|_2^2 \leq M^2.$$
- Smoothness or convexity of f.

Two evaluation choices:

- Averaged iterate: $\bar{x}_T := \frac{1}{T+1} \sum_{k=0}^{T} x_k$ (easier bounds).
- Last iterate: x_T (more relevant in practice; harder analysis).

Convex Case

Convergence of SGD for convex functions

Assume f is convex and $\sup_x \mathbb{E}\,\|g(x,\xi)\|_2^2 \le M^2$. Let $R \ge \|x_0 - x^*\|_2$, choose a constant step size

$$\eta_k \equiv \eta := \frac{R}{M\sqrt{T+1}}\,.$$

Then for the averaged iterate $\bar{x}_T$:

$$\mathbb{E}[f(\bar{x}_T) - f(x^*)] \le \frac{RM}{\sqrt{T+1}}.$$

Rate: in nonsmooth convex problems, SGD matches the optimal $\Theta(1/\sqrt{T})$ rate (up to constants).

Proof Sketch

Start from the squared distance recursion:

$$\begin{aligned}\mathbb{E}\left\|x_{k+1}-x^*\right\|_2^2 &= \mathbb{E}\left\|x_k-\eta g(x_k,\xi_k)-x^*\right\|_2^2 \\ &= \mathbb{E}\left\|x_k-x^*\right\|_2^2+\eta^2\mathbb{E}\left\|g(x_k,\xi_k)\right\|_2^2 \\ &\quad -2\eta\,\mathbb{E}\left\langle g(x_k,\xi_k),x_k-x^*\right\rangle\end{aligned}$$

Plug in assumptions:

- Unbiasedness: $\mathbb{E}\left\langle g(x_k,\xi_k),x_k-x^*\right\rangle = \mathbb{E}\left\langle \nabla f(x_k),x_k-x^*\right\rangle$.
- Convexity: $\left\langle \nabla f(x_k),x_k-x^*\right\rangle \geq f(x_k)-f(x^*)$.
- Second moment: $\mathbb{E}\left\|g(x_k,\xi_k)\right\|_2^2 \leq M^2$.

Resulting inequality:

$$2\eta\,\mathbb{E}[f(x_k)-f(x^*)] \leq \mathbb{E}\left\|x_k-x^*\right\|_2^2-\mathbb{E}\left\|x_{k+1}-x^*\right\|_2^2+\eta^2M^2.$$

Telescoping $k=0,\ldots,T$ **and Jensen on** $\bar{x}_T \implies$ stated bound.

Strongly Convex Case

Convergence of SGD for strongly convex functions

Assume f is μ-strongly convex and $\sup_x \mathbb{E}\,\|g(x,\xi)\|_2^2 \le M^2$. Choose diminishing step sizes

$$\eta_k = \frac{1}{\mu(k+1)}.$$

Then

$$\mathbb{E}[f(\bar{x}_T) - f(x^*)] \le \frac{M^2(1+\log(T+1))}{2\mu(T+1)}, \quad \mathbb{E}\,\|x_k - x^*\|_2^2 \le \frac{Q}{k+1},$$

where $Q := \max\{M^2/\mu^2, \|x_0 - x^*\|_2^2\}$.

Rate: compared to GD, SGD has an extra $\log T$ factor.

Proof Sketch

Strong convexity:

$$f(x_k) - f(x^*) \leq \langle \nabla f(x_k), x_k - x^* \rangle - \frac{\mu}{2} \|x_k - x^*\|_2^2 .$$

Dsistance recursion:

$$\mathbb{E}[f(x_k) - f(x^*)] \leq \frac{\mathbb{E}\|x_k - x^*\|_2^2 - \mathbb{E}\|x_{k+1} - x^*\|_2^2}{2\eta_k} + \frac{\eta_k M^2}{2} - \frac{\mu}{2}\mathbb{E}\|x_k - x^*\|_2^2 .$$

Pick $\eta_k = 1/(\mu(k+1))$**:**

- objective gap bound comes from summing η_k (harmonic series);
- distance bound follows from a contraction-style recursion.

Practical note: averaging ($\bar{x}_T$) is a common theoretical trick; last-iterate bounds are more subtle.

Last-Iterate Convergence

Issue: bounds for $\bar{x}_T$ do not directly explain why the last iterate often works well.

The following last-iterate result is from [Shamir and Zhang, 2013].

Last-iterate convergence of SGD

Under μ-strong convexity and $\sup_x \mathbb{E}\|g(x,\xi)\|_2^2 \le M^2$, using

$$\eta_k = \frac{1}{\mu k},$$

one can show

$$\mathbb{E}[f(x_T) - f(x^*)] \le \frac{17M^2(1+\log T)}{\mu T}.$$

Takeaway: last-iterate can match averaged-iterate rates (up to constants), but proofs require refined arguments.

Constraints and Projection

Constrained stochastic optimization:

$$\min_{x \in \mathcal{X}} f(x) = \mathbb{E}_{\xi \sim \mathcal{D}}[F(x, \xi)], \quad \mathcal{X} \subset \mathbb{R}^d \text{ convex.}$$

Projected SGD: $x_{k+1} = \Pi_{\mathcal{X}}(x_k - \eta_k \, g(x_k, \xi_k))$.

The following last-iterate result is from [Shamir and Zhang, 2013].

Last-iterate convergence of projected SGD

Assume f is convex and

$$\sup_x \mathbb{E}_{\xi \sim \mathcal{D}}\left[\|g(x, \xi)\|_2^2\right] \le M^2, \qquad \sup_{x, x' \in \mathcal{X}} \|x - x'\|_2 \le D.$$

If $\eta_k = \frac{c}{\sqrt{k}}$ for some $c > 0$, then for any $T > 1$,

$$\mathbb{E}[f(x_T) - f(x^*)] \le \left(\frac{D^2}{c} + cM^2\right) \frac{2 + \log(T)}{\sqrt{T}}.$$

Tuning: choosing $c \approx D/M$ balances the two terms.

High-Probability Guarantees

Why high probability? In practice, we would like stronger guarantees since it is not practical to run these algorithms for a long run and consider the average error.

The following last-iterate result is from [Chou et al., 2020].

Last-iterate high-probability convergence of SGD

Assume μ-strong convexity and $\sup_x \mathbb{E}_{\xi\sim D}[g(x,\xi)^2] \le M$. With $\eta_k = \frac{1}{\mu(k+1)}$, one can prove that for $\delta \in (0,1)$,

$$\mathbb{P}\left\{\|x_T - x^*\|_2^2 \ge \frac{500M^2\log(1/\delta)}{\mu^2 T}\right\} \le \delta.$$

Furthermore, for any $1 \le t \le T$,

$$\mathbb{P}\left\{\|x_t - x^*\|_2^2 \ge \frac{1000M^2(\log(1/\delta)+\log(\log(t+1)))}{\mu^2 t}\right\} \le \delta$$

Interpretation: with probability $\ge 1-\delta$, the last iterate is within a radius $O\left(\sqrt{\frac{\log(1/\delta)}{T}}\right)$ of x^*.

Outline

Stochastic Optimization

Convergence of SGD

Variance Reduction

Variance Reduction

Objective:

$$\min_{x \in \mathbb{R}^d} \; f(x) := \frac{1}{n} \sum_{i=1}^{n} f_i(x).$$

Assumptions:

- each f_i is ℓ-smooth: $\|\nabla f_i(x) - \nabla f_i(y)\|_2 \leq \ell \|x - y\|_2$;
- f is μ-strongly convex; condition number $\kappa := \frac{\ell}{\mu}$;
- computing $\nabla f_i(x)$ costs $O(1)$, while $\nabla f(x)$ costs $O(n)$.

Goal: keep $O(1)$ per-step cost and recover fast (GD-like) convergence.

GD vs. SGD: What is the Gap?

Method	Gradient evals	Typical total gradients
GD	n	$O\left(n\kappa \log \frac{1}{\varepsilon}\right)$
SGD	1	$O\left(\frac{1}{\mu\varepsilon}\right)$

SGD under strongly convex, with $\eta_k = \frac{1}{\mu(k+1)}$, $\mathbb{E}\|x_k - x^*\|_2^2 = O(1/k)$.
However, even if we add ℓ-smoothness, $\mathbb{E}[f(x_k) - f(x^*)] = O(1/k)$.

Bottleneck: SGD gradients are unbiased but have non-vanishing variance $\implies$ need diminishing η_k and cannot recover linear convergence in GD.

Variance reduction: build an unbiased estimator whose variance shrinks as we approach x^*.

Constant Step SGD

Example: $f(x) = \frac{1}{2}x^2$, so $\mu = L = 1$, and minimizer $x^* = 0$.

Stochastic gradient: $g(x) = x + \zeta$, where $\mathbb{E}[\zeta] = 0$ and $\mathbb{E}[\zeta^2] = \sigma^2 > 0$.

SGD with constant step α:

$$x_{k+1} = x_k - \alpha g(x_k) = (1-\alpha)x_k - \alpha\zeta_k.$$

Square and take expectation:

$$\mathbb{E}[x_{k+1}^2] = (1-\alpha)^2\mathbb{E}[x_k^2] + \alpha^2\sigma^2.$$

Solve the recursion:

$$\mathbb{E}[x_k^2] = (1-\alpha)^{2k}x_0^2 + \frac{\alpha^2\sigma^2}{1-(1-\alpha)^2} = (1-\alpha)^{2k}x_0^2 + \frac{\alpha\sigma^2}{2-\alpha}.$$

Conclusion: not convergent to minimizer

$$\lim_{k\to\infty} \mathbb{E}[x_k^2] = \frac{\alpha\sigma^2}{2-\alpha} \approx \frac{\alpha\sigma^2}{2} > 0.$$

Control Variate

Fix a snapshot $\tilde{x}$ and compute $\nabla f(\tilde{x})$ once. Sample $i \sim \mathrm{Unif}\{1, \ldots, n\}$.

Control variate: let $X := \nabla f_i(x)$ and $Y := \nabla f_i(\tilde{x})$, so

$$\mathbb{E}[X \mid x] = \nabla f(x), \qquad \mathbb{E}[Y \mid \tilde{x}] = \nabla f(\tilde{x}).$$

Define the SVRG estimator

$$v(x; \tilde{x}, i) := X - (Y - \mathbb{E}[Y]) = \nabla f_i(x) - \nabla f_i(\tilde{x}) + \nabla f(\tilde{x}),$$

so $\mathbb{E}[v \mid x, \tilde{x}] = \nabla f(x)$.

Variance we shrink: the estimation noise

$$v - \nabla f(x) = (X - \mathbb{E}X) - (Y - \mathbb{E}Y).$$

If $x \approx \tilde{x}$, then X and Y are highly correlated and noise cancels; moreover if f_i is ℓ-smooth,

$$\mathbb{E}\|v - \nabla f(x)\|_2^2 \leq \mathbb{E}\|X - Y\|_2^2 \leq \ell^2 \|x - \tilde{x}\|_2^2.$$

SVRG

Goal: minimize $f(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x)$.

Outer loop (epoch) $s = 0, 1, \ldots, S-1$**:**

- Snapshot: $\tilde{x} \leftarrow \tilde{x}^s$
- Full gradient: $\mu \leftarrow \nabla f(\tilde{x}) = \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(\tilde{x})$
- Initialize: $x_0 \leftarrow \tilde{x}$
- Choose inner length $N \in \{0, \ldots, m-1\}$ uniformly (or set $N = m$)
- Inner loop $k = 0, 1, \ldots, N-1$:
  - Sample $i_k \sim \mathrm{Unif}\{1, \ldots, n\}$
  - Form $v_k = \nabla f_{i_k}(x_k) - \nabla f_{i_k}(\tilde{x}) + \mu$
  - Update $x_{k+1} = x_k - \eta\, v_k$
- Epoch update: $\tilde{x}^{s+1} \leftarrow x_N$

Work per epoch: one full gradient (n) + N $(\approx m)$ stochastic gradients.

SVRG was proposed in [Johnson and Zhang, 2013].

SVRG Convergence

Convergence of SVRG

Assume each f_i is ℓ-smooth and f is μ-strongly convex. Choose $\eta = \frac{1}{8\ell}$ and $m \geq 16\kappa$ where $\kappa = \frac{\ell}{\mu}$. Then the snapshot points satisfy

$$\mathbb{E}[f(\tilde{x}^s) - f(x^*)] + \mu\,\mathbb{E}\left[\|\tilde{x}^s - x^*\|_2^2\right] \leq \left(\tfrac{1}{2}\right)^s \left(f(\tilde{x}^0) - f(x^*) + \mu\|\tilde{x}^0 - x^*\|_2^2\right).$$

Complexity: each epoch costs $(n+m)$ gradients, and $m = \Theta(\kappa)$, so total gradient complexity is $O\left((n+\kappa)\log\frac{1}{\varepsilon}\right)$.

Method	Gradient evals	Typical total gradients
GD	n	$O\left(n\kappa \log \frac{1}{\varepsilon}\right)$
SGD	1	$O\left(\frac{1}{\mu\varepsilon}\right)$
SVRG	$n+m,\ m = \Theta(\kappa)$	$O\left((n+\kappa)\log\frac{1}{\varepsilon}\right)$

Proof Sketch

Let $\Delta(z) := f(z) - f(x^*)$ and $v = \nabla f_i(x) - \nabla f_i(\tilde{x}) + \nabla f(\tilde{x})$.

Variance bound: $\mathbb{E}_i[v \mid x, \tilde{x}] = \nabla f(x)$ and

$$\mathbb{E}_i\|v - \nabla f(x)\|_2^2 \le \mathbb{E}_i\|\nabla f_i(x) - \nabla f_i(\tilde{x})\|_2^2 \le 4\ell(\Delta(x) + \Delta(\tilde{x})).$$

The last step uses $\|\nabla f_i(u) - \nabla f_i(x^*)\|_2^2 \le 2\ell(f_i(u) - f_i(x^*))$.

Bound $\mathbb{E}\|v\|_2^2$. Since $\mathbb{E}_i[v \mid x, \tilde{x}] = \nabla f(x)$,

$$\mathbb{E}_i\|v\|_2^2 = \|\nabla f(x)\|_2^2 + \mathbb{E}_i\|v - \nabla f(x)\|_2^2 \le 6\ell\Delta(x) + 4\ell\Delta(\tilde{x}).$$

Telescoping. Let $d := x - x^*$. Then $d^+ = x^+ - x^* = d - \eta v$ and

$$\mathbb{E}_i\|d^+\|_2^2 = \|d\|_2^2 - 2\eta\langle\nabla f(x), d\rangle + \eta^2\mathbb{E}_i\|v\|_2^2.$$

Strong convexity gives $\langle\nabla f(x), d\rangle \ge \Delta(x) + \frac{\mu}{2}\|d\|_2^2$, hence

$$\mathbb{E}_i\|d^+\|_2^2 \le (1 - \eta\mu)\|d\|_2^2 - 2\eta(1 - 3\ell\eta)\Delta(x) + 4\ell\eta^2\Delta(\tilde{x}).$$

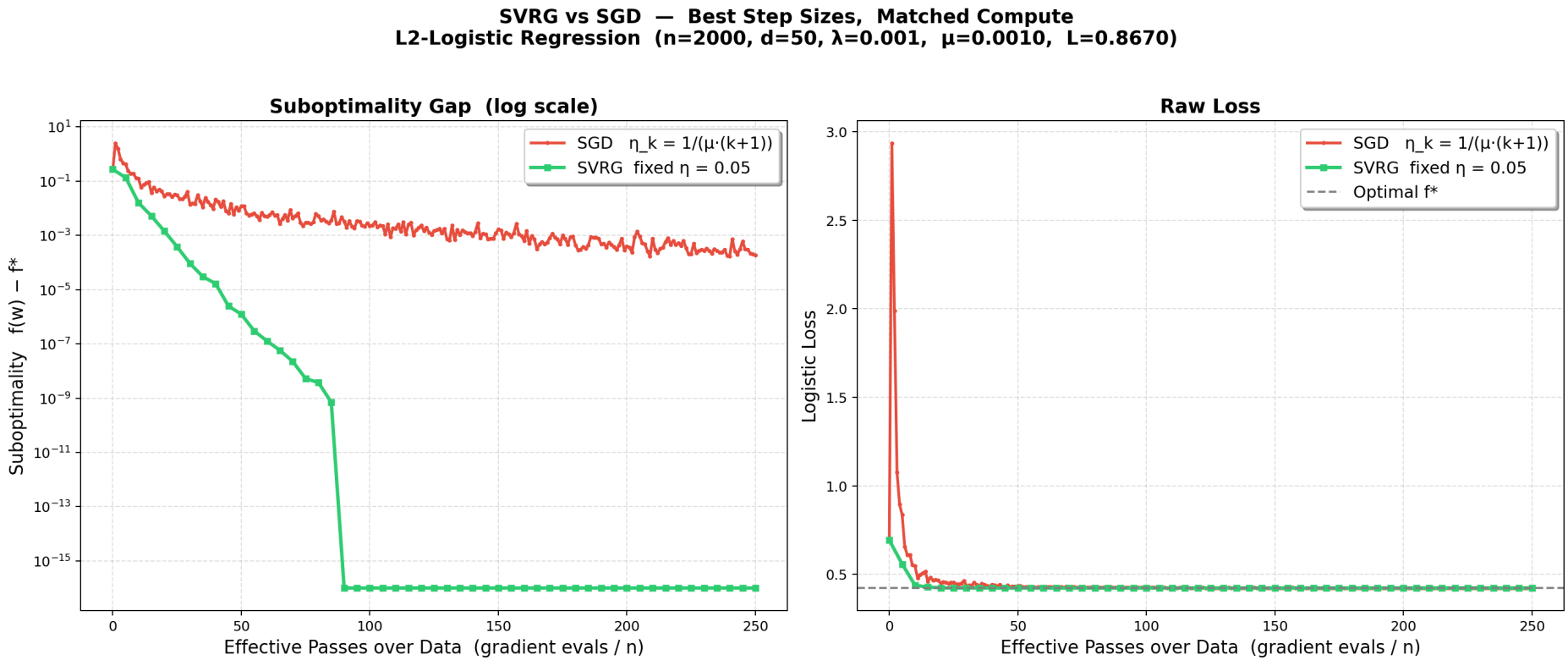
Choosing $\eta = \frac{1}{8\ell}$ makes $2(1 - 3\ell\eta) = \frac{5}{4} > 1$ and yields a one-epoch contraction after summing over $m = \Theta(\kappa)$ inner steps.

Numerical Example

Problem: ℓ_2-regularised logistic regression

$$\min_w \; f(w) = \frac{1}{n}\sum_{i=1}^{n} \log\!\left(1 + e^{-y_i x_i^\top w}\right) + \frac{\lambda}{2}\|w\|^2$$

with $n = 2000$, $d = 50$, $\lambda = 10^{-3}$, $\mu \approx 10^{-3}$, $L \approx 0.87$, $\kappa = L/\mu \approx 867$.



Alternatives: SAG & SAGA

Setting: $f(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x)$.

Key idea: store a memory of per-sample gradients and update one entry each step. Maintain $\{y_i\}_{i=1}^{n}$ and their average $\bar{y} := \frac{1}{n}\sum_{i=1}^{n} y_i$.

At iteration t: sample $i_t \sim \mathrm{Unif}\{1, \dots, n\}$ and compute $g_t = \nabla f_{i_t}(x_t)$. Update memory:

$$y_{i_t} \leftarrow g_t, \quad \bar{y} \leftarrow \bar{y} + \frac{1}{n}(g_t - y_{i_t}^{\text{old}}).$$

SAG update (biased):

$$x_{t+1} = x_t - \eta\, \bar{y}.$$

SAGA update (unbiased):

$$x_{t+1} = x_t - \eta(g_t - y_{i_t}^{\text{old}} + \bar{y}), \quad \mathbb{E}[g_t - y_{i_t}^{\text{old}} + \bar{y} \mid x_t] = \nabla f(x_t).$$

Takeaway: variance shrinks as y_i's become consistent near x^*.

Discussion

SVRG:

- **Pros:** small memory; strong theory under strong convexity.
- **Cons:** needs periodic full gradient; double loop, harder to tune; no theory under non-strong convexity or even non-convexity.

SAG/SAGA:

- **Pros:** no full-gradient computation; single loop, easier to tune; strong theory beyond strong convexity.
- **Cons:** needs storing n gradients. SAG estimator is biased.

Lesson:

- **Memory tight:** prefer SVRG.
- **Full-gradient not available:** prefer SAGA or SAG.
- **Very large-scale:** consider mini-batch VR variants (e.g., SARAH/SPIDER) or adaptive methods.

Wrap-up and Next Week

Week 6: Key takeaways

- Stochastic optimization, population and finite-sum form.
- SGD with different sampling techniques.
- Convergence of SGD for average-iterate, last-iterate, in expectation and with high probability.
- Variance reduction, SVRG, complexity $O((n+\kappa)\log(1/\varepsilon))$.

Thank you for your attendance!

Questions?

Next week (Week 7): SGDM & Adam-family

References

Chou, C.-N., Sandhu, J. S., Wang, M. B., and Yu, T. (2020).
A general framework for analyzing stochastic dynamics in learning algorithms.
arXiv preprint arXiv:2006.06171.

Johnson, R. and Zhang, T. (2013).
Accelerating stochastic gradient descent using predictive variance reduction.
Advances in neural information processing systems, 26.

Shamir, O. and Zhang, T. (2013).
Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes.
In *Proceedings of the 30th International Conference on Machine Learning (ICML)*.