# A Theory of Variance Reduction for Optimization and Sampling based on Importance Sampling

Ayoub El Hanchi

December 18, 2020

**Abstract**

Most variance-reduced algorithms for finite-sum optimization and sampling problems rely on the use of control variates to reduce the variance of the gradient estimator. An alternative way of achieving variance reduction is through importance sampling. There is a large literature on such methods but until recently most proposed algorithms were based on heuristics, and came with little to no theoretical guarantees. In this paper, we build a theory of variance-reduced algorithms based on importance sampling. We develop two new procedures that we refer to as stochastic reweighted gradient (SRG) and stochastic reweighted gradient Langevin dynamics (SRG-LD). We provide guarantees for both SRG and SRG-LD using both constant and decreasing step-sizes. We show that they can provably outperform SGD and SGLD while requiring the exact same number of gradient evaluations and introducing negligible memory and time overheads.

## 1 Introduction

We consider finite-sum problems of the form:

$$F(x) \coloneqq \sum_{i=1}^{n} f_i(x) \tag{1}$$

where $F : \mathbb{R}^d \to \mathbb{R}$ is the objective in optimization, or the log-density in sampling. Problems of this form are common in machine learning and statistics. When $n$ is large, computing the gradient of $F$ becomes prohibitively expensive, and stochastic estimates are used instead. Stochastic gradient descent (SGD) and stochastic gradient Langevin dynamics (SGLD) are the most basic stochastic algorithms for optimization and sampling respectively. To evaluate the gradient at a given point $x$, they employ the simple estimator $n\nabla f_j(x)$ where $j$ is uniformly distributed over $[n]$. While this estimator is unbiased, it can suffers from high-variance, which can severely affect the convergence of the algorithms.

A long line of work has led to a new set of methods that are able to provably reduce the variance of this gradient estimator and achieve fast convergence rates. These variance-reduced methods rely on the construction of control variates, most commonly by storing previously seen gradients, or by periodically computing the full gradient. They can be shown to be optimal as they achieve known lower-bounds in the gradient oracle model of complexity. These methods however suffer from two drawbacks. On the one hand, just like SGD and SGLD, they require random access to the gradient of individual functions $f_i$, which can cause significant slowdown, particularly in the nowadays common case where the gradient computation and the storage of the data are done on

separate devices. On the other, they either require a prohibitively large amount of memory, or the periodic computation of the full gradient which can make per-iteration progress unacceptably slow.

Another generic method used to reduce the variance of a given Monte Carlo estimator is importance sampling. While there is a large literature on importance sampling methods for SGD, most proposed methods rely on heuristics and come with little to no theoretical convergence guarantees. A recent trend in this literature relied on an online formulation of the problem of designing adaptive probabilities to show that certain methods are able to achieve sublinear regret, giving some justification for their use. Directly proving improved convergence guarantees on the optimization or sampling algorithms themselves has remained elusive however.

In this paper, we propose two new variance-reduced algorithms based on importance sampling: stochastic reweighted gradient (SRG) for optimization and stochastic reweighted Langevin dynamics (SRG-LD) for sampling.

# 2 Algorithms

**Algorithm 1:** Stochastic Reweighted Gradient (SRG)

**Parameters:** step sizes $(\alpha_k)_{k=1}^{\infty} > 0$, lower bounds $(\varepsilon_k)_{k=1}^{\infty} \in (0, \frac{1}{n}]$
**Initialization:** $x_0 \in \mathbb{R}^d, (g_0^i)_{i=1}^n \in \mathbb{R}^d$
**for** $k = 0, 1, 2, \ldots$ **do**
  $p_k = \arg\min_{p \in \Delta(\varepsilon_k)} \sum_{i=1}^n \frac{1}{p^i} \left\| g_k^i \right\|_2^2$
  sample $i_k \sim p_k$
  $x_{k+1} = x_k - \alpha_k \frac{1}{p_k^{i_k}} \nabla f_{i_k}(x_k)$
  sample $b_k \sim \text{Bernoulli}(\frac{\varepsilon_k}{p_k^{i_k}})$
  $g_{k+1}^i = \begin{cases} \nabla f_i(x_k) & \text{if } i = i_k \text{ and } b_k = 1 \\ g_k^i & \text{otherwise} \end{cases}$
**end**

# 3 Convergence Analysis

## 3.1 Assumptions

**Assumption 1.** *The function $F : \mathbb{R}^d \to \mathbb{R}$ is differentiable and $\mu$-strongly convex, that is:*

$$F(y) \geq F(x) + \langle \nabla F(x), y - x \rangle + \frac{\mu}{2} \left\| y - x \right\|_2^2 \quad \forall x, y \in \mathbb{R}^d$$

**Assumption 2.** *For all $i \in [n]$, the functions $f_i : \mathbb{R}^d \to \mathbb{R}$ are differentiable and convex, that is:*

$$f_i(y) \geq f_i(x) + \langle \nabla f_i(x), y - x \rangle \quad \forall x, y \in \mathbb{R}^d$$

**Assumption 3.** *For all $i \in [n]$, the functions $f_i : \mathbb{R}^d \to \mathbb{R}$ are $L$-smooth, that is:*

$$f_i(y) \leq f_i(x) + \langle \nabla f_i(x), y - x \rangle + \frac{L}{2} \left\| y - x \right\|_2^2 \quad \forall x, y \in \mathbb{R}^d$$

**Lemma 1.** *For all $i \in [n]$ we have:*

$$f_i(y) \geq f_i(x) + \langle \nabla f_i(x), y - x \rangle + \frac{1}{2L} \|\nabla f_i(y) - \nabla f_i(x)\|_2^2$$

### 3.2 Bounding sub-optimality of probabilities

**Lemma 2.** *Let $\{a_i\}_{i=1}^n$ be a non-negative set of numbers where at least one of the $a_i$'s is strictly positive. Then:*

$$\arg\min_{p \in \Delta(\varepsilon)} \sum_{i=1}^n \frac{1}{p^i} a_i^2 \leq (1 + 2n\varepsilon) \arg\min_{p \in \Delta} \sum_{i=1}^n \frac{1}{p^i} a_i^2 = (1 + 2n\varepsilon) \left( \sum_{i=1}^n a_i \right)^2$$

*for all $0 \leq \varepsilon \leq \frac{1}{2n}$.*

### 3.3 Useful lemmas

**Lemma 3.** *Let $k \in \mathbb{N}$. Taking expectation with respect to $i_k$ and $b_k$ conditional on $(i_t)_{t=0}^{k-1}$ and $(b_t)_{t=0}^{k-1}$, we have:*

$$\mathbb{E}\left[ \sum_{i=1}^n \left\| g_{k+1}^i - \nabla f_i(x^*) \right\|_2^2 \right] \leq (1 - \varepsilon_k) \sum_{i=1}^n \left\| g_k^i - \nabla f_i(x^*) \right\| + 2L\varepsilon_k \left[ F(x_k) - F(x^*) \right]$$

*Proof.*

$$\begin{aligned}
&\mathbb{E}\left[ \sum_{i=1}^n \left\| g_{k+1}^i - \nabla f_i(x^*) \right\|_2^2 \right] \\
&= \sum_{j=1}^n \mathbb{P}(i_k = j) \Bigg[ \mathbb{P}(b_k = 0 \mid i_k = j) \left( \sum_{i=1}^n \left\| g_k^i - \nabla f_i(x^*) \right\|_2^2 \right) \\
&\quad + \mathbb{P}(b_k = 1 \mid i_k = j) \left( \sum_{\substack{i=1 \\ i \neq j}}^n \left\| g_k^i - \nabla f_i(x^*) \right\|_2^2 + \|\nabla f_j(x_k) - \nabla f_j(x^*)\|_2^2 \right) \Bigg] \\
&= \sum_{j=1}^n p_k^j \left[ \left( 1 - \frac{\varepsilon_k}{p_k^j} \right) \sum_{i=1}^n \left\| g_k^i - \nabla f_i(x^*) \right\|_2^2 + \frac{\varepsilon_k}{p_k^j} \left( \sum_{\substack{i=1 \\ i \neq j}}^n \left\| g_k^i - \nabla f_i(x^*) \right\|_2^2 + \|\nabla f_j(x_k) - \nabla f_j(x^*)\|_2^2 \right) \right] \\
&= (1 - \varepsilon_k) \sum_{i=1}^n \left\| g_k^i - \nabla f_i(x^*) \right\|_2^2 + \varepsilon_k \sum_{i=1}^n \|\nabla f_i(x_k) - \nabla f_i(x^*)\|_2^2 \\
&\leq (1 - \varepsilon_k) \sum_{i=1}^n \left\| g_k^i - \nabla f_i(x^*) \right\|_2^2 + 2L\varepsilon_k \left[ F(x_k) - F(x^*) \right]
\end{aligned}$$

□

**Lemma 4.** *For all $\beta, \gamma, \delta, \eta > 0$ we have:*

$$\begin{aligned}\sum_{i=1}^{n} \frac{1}{p_k^i} \|\nabla f_i(x_k)\|_2^2 \leq (1+\beta+\gamma)\frac{2L}{\varepsilon_k}\left[F(x_k) - F^*\right] \\ + \left((1+\beta^{-1}+\delta) + (1+\gamma^{-1}+\delta^{-1})(1+\eta)\right)\frac{1}{\varepsilon_k}\sum_{i=1}^{n}\left\|g_k^i - \nabla f_i(x^*)\right\|_2^2 \\ + (1+\gamma^{-1}+\delta^{-1})(1+\eta^{-1})\sigma_*^2\end{aligned}$$

*where:*

$$\sigma_*^2 \coloneqq \arg\min_{p\in\Delta} \sum_{i=1}^{n} \frac{1}{p^i}\|\nabla f_i(x^*)\|_2^2 = \left(\sum_{i=1}^{n}\|\nabla f_i(x^*)\|_2\right)^2$$

*Proof.*

$$\begin{aligned}\sum_{i=1}^{n} \frac{1}{p_k^i} \|\nabla f_i(x_k)\|_2^2 &= \sum_{i=1}^{n} \frac{1}{p_k^i} \left\|\nabla f_i(x_k) - \nabla f_i(x^*) + \nabla f_i(x^*) - g_k^i + g_k^i\right\|_2^2 \\ &\leq (1+\beta+\gamma)\sum_{i=1}^{n}\frac{1}{p_k^i}\|\nabla f_i(x_k) - \nabla f_i(x^*)\|_2^2 \\ &\quad + (1+\beta^{-1}+\delta)\sum_{i=1}^{n}\frac{1}{p_k^i}\left\|g_k^i - \nabla f_i(x^*)\right\|_2^2 \\ &\quad + (1+\gamma^{-1}+\delta^{-1})\sum_{i=1}^{n}\frac{1}{p_k^i}\left\|g_k^i\right\|_2^2\end{aligned}$$

Let us bound each of the three terms. The first is bound by:

$$\begin{aligned}\sum_{i=1}^{n}\frac{1}{p_k^i}\|\nabla f_i(x_k) - \nabla f_i(x^*)\|_2^2 &\leq \frac{1}{\varepsilon_k}\sum_{i=1}^{n}\|\nabla f_i(x_k) - \nabla f_i(x^*)\|_2^2 \\ &\leq \frac{2L}{\varepsilon_k}\left[F(x_k) - F(x^*)\right]\end{aligned}$$

The second is easily bound by:

$$\sum_{i=1}^{n}\frac{1}{p_k^i}\left\|g_k^i - \nabla f_i(x^*)\right\|_2^2 \leq \frac{1}{\varepsilon_k}\sum_{i=1}^{n}\left\|g_k^i - \nabla f_i(x^*)\right\|_2^2$$

Finally, the third term can be bound as:

$$
\begin{aligned}
&\sum_{i=1}^{n} \frac{1}{p_k^i} \left\|g_k^i\right\|_2^2 \\
&\leq (1+2n\varepsilon_k)\left(\sum_{i=1}^{n} \left\|g_k^i\right\|_2\right) \\
&\leq (1+2n\varepsilon_k)\left(\sum_{i=1}^{n} \left\|g_k^i - \nabla f_i(x^*)\right\|_2 + \sum_{i=1}^{n} \left\|\nabla f_i(x^*)\right\|_2\right) \\
&\leq (1+2n\varepsilon_k)(1+\eta)\left(\sum_{i=1}^{n} \left\|g_k^i - \nabla f_i(x_k)\right\|_2\right)^2 + (1+2n\varepsilon_k)(1+\eta^{-1})\left(\sum_{i=1}^{n} \left\|\nabla f_i(x^*)\right\|_2\right)^2 \\
&\leq (1+2n\varepsilon_k)(1+\eta)n\sum_{i=1}^{n} \left\|g_k^i - \nabla f_i(x^*)\right\|_2^2 + (1+2n\varepsilon_k)(1+\eta^{-1})\sigma_*^2 \\
&\leq 2n(1+\eta)\sum_{i=1}^{n} \left\|g_k^i - \nabla f_i(x^*)\right\|_2^2 + (1+2n\varepsilon_k)(1+\eta^{-1})\sigma_*^2 \\
&\leq (1+\eta)\frac{1}{\varepsilon_k}\sum_{i=1}^{n} \left\|g_k^i - \nabla f_i(x^*)\right\|_2^2 + (1+\eta^{-1})(1+2n\varepsilon_k)\sigma_*^2
\end{aligned}
$$

Combining the bounds yields the result. □

### 3.4 Main Theorem

**Theorem 1.** *Let $x^*$ be the unique minimizer of $F$, and suppose that Assumptions 1, 2 and 3 hold. Further, assume that for all $k \in \mathbb{N}$:*

- $\frac{\alpha_k}{\varepsilon_k}$ *is constant.*
- $\varepsilon_k \leq \frac{1}{2n}$.
- $\alpha_k \leq \frac{1}{20}\frac{\varepsilon_k}{L}$.

*Define the Lyapunov function:*

$$
T^k \coloneqq T(x_k, (g_k^i)_{i=1}^n) \coloneqq \frac{\alpha_k}{\varepsilon_k}\frac{a}{L}\sum_{i=1}^{n} \left\|g_k^i - \nabla f_i(x^*)\right\|_2^2 + \left\|x_k - x^*\right\|_2^2
$$

*with $a \coloneqq 0.673$ and where $(x_k, (g_k^i)_{i=1}^n)$ evolves according to Algorithm 1. Then for all $k \in \mathbb{N}$:*

$$
\mathbb{E}\left[T^{k+1}\right] \leq (1-\alpha_k\rho)T^k + (1+2n\varepsilon_k)2\alpha_k^2\sigma_*^2
$$

*where $\rho \coloneqq \min\{L, \mu\}$.*

*Proof.* All the expectations in this proof are conditional on $(i_t)_{t=0}^{k-1}$ and $(b_t)_{t=0}^{k-1}$ and are taken over

$i_k$ and $b_k$. We have:

$$\begin{aligned}
\mathbb{E}\left[\|x_{k+1}-x^*\|_2^2\right] &= \mathbb{E}\left[\left\|x_k-\alpha_k\frac{1}{p_k^{i_k}}\nabla f_{i_k}(x_k)-x^*\right\|_2^2\right] \\
&= \|x_k-x^*\|_2^2 - 2\alpha_k\langle\mathbb{E}\left[\frac{1}{p_k^{i_k}}\nabla f_1 i_k(x_k)\right], x_k-x^*\rangle + \alpha_k^2\mathbb{E}\left[\left\|\frac{1}{p_k^{i_k}}\nabla f_{i_k}(x_k)\right\|_2^2\right] \\
&= \|x_k-x^*\|_2^2 - 2\alpha_k\langle\nabla F(x_k), x_k-x^*\rangle + \alpha_k^2\sum_{i=1}^n\frac{1}{p_k^i}\|\nabla f_i(x_k)\|_2^2 \\
&\leq (1-\alpha_k\mu)\|x_k-x^*\|_2^2 - 2\alpha_k\left[F(x_k)-F(x^*)\right] + \alpha_k^2\sum_{i=1}^n\frac{1}{p_k^i}\|\nabla f_i(x_k)\|_2^2
\end{aligned}$$

Bounding the last term using Lemma 4, and combining the obtained bound with the bound on the first term of the Lyapunov function $T^{k+1}$ from Lemma 3 we get the overall bound:

$$\begin{aligned}
\mathbb{E}\left[T^{k+1}\right] \leq &\left(1-\varepsilon_k+\frac{D_1\alpha_k L}{a}\right)\frac{\alpha_k}{\varepsilon_k}\frac{a}{L}\sum_{i=1}^n\left\|g_k^i-\nabla f_i(x^*)\right\|_2^2 \\
&+ (1-\alpha_k\mu)\|x_k-x^*\|_2^2 \\
&+ \frac{2\alpha_k L}{\varepsilon_k}\left(D_2\alpha_k+\frac{a\varepsilon_k}{L}-\frac{\varepsilon_k}{L}\right)\left[F(x_k)-F(x^*)\right] \\
&+ D_3(1+2n\varepsilon_k)\alpha_k^2\sigma_*^2
\end{aligned}$$

where:

$$\begin{aligned}
D_1 &\coloneqq (1+\beta^{-1}+\delta)+(1+\gamma^{-1}+\delta^{-1})(1+\eta) \\
D_2 &\coloneqq (1+\beta+\gamma) \\
D_3 &\coloneqq (1+\gamma^{-1}+\delta^{-1})(1+\eta^{-1})
\end{aligned}$$

To ensure that the third term is less than zero we need:

$$\alpha_k \leq \frac{(1-a)}{D_2}\frac{\varepsilon_k}{L} \tag{2}$$

Replacing in the first parentheses we get:

$$1-\varepsilon_k+\frac{D_1\alpha_k L}{a} \leq 1-\alpha_k\left(\frac{D_2}{(1-a)}-\frac{D_1}{a}\right)L$$

the final bound, assuming that $\alpha_k$ satisfies (2), is given by:

$$\mathbb{E}\left[T^{k+1}\right] \leq (1-\alpha_k\xi)\,T^k + D_3(1+2n\varepsilon_k)\alpha_k^2\sigma_*^2$$

where:

$$\xi \coloneqq \min\left\{\left(\frac{D_2}{1-a}-\frac{D_1}{a}\right)L, \mu\right\}$$

Our goal now is to choose the parameters $a, \beta, \gamma, \delta, \eta$ so as to minimize the bound. Before doing so we constrain our problem in the following ways. First, to allow the comparison with the uniform

sampling case, we choose to enforce $D_3 \leq 2$. We then would ideally like to maximize $\alpha_k \xi$ subject to (2) directly. However, we don't know a priori the values of $L$ and $\mu$, so we first use the bound:

$$\xi \leq \min\left\{\frac{D_2}{1-a} - \frac{D_1}{a}, 1\right\}\rho$$

where $\rho \coloneqq \min\{L, \mu\}$. Maximizing the upper bound can now be shown to be equivalent to solving the constrained optimization problem:

$$\begin{aligned} \max_{a,\beta,\gamma,\delta,\eta} \quad & \frac{(1-a)}{D_2} \\ \text{subject to:} \quad & \frac{D_2}{1-a} - \frac{D_1}{a} \geq 1 \\ & D_3 \leq 2 \\ & 0 < a < 1 \\ & \beta, \gamma, \delta, \eta > 0 \end{aligned}$$

We solve this problem numerically to find a feasible point which closely approximates the solution:

$$a = 0.673,\ \beta = 1.028,\ \gamma = 4.084,\ \delta = 3.973,\ \eta = 2.980$$

With this choice of free parameters we get our final bound:

$$\mathbb{E}\left[T^{k+1}\right] \leq (1 - \alpha_k \rho) T^k + 2(1 + 2n\varepsilon_k)\alpha_k^2 \sigma_*^2$$

with the assumption that the following conditions hold for all $k \in \mathbb{N}$:

- $\varepsilon_k \leq \frac{1}{2n}$.
- $\alpha_k \leq \frac{1}{20}\frac{\varepsilon_k}{L}$.
- $\frac{\alpha_k}{\varepsilon_k}$ is constant.

□

**Corollary 1.** *Assume that Assumptions 1, 2 and 3 hold, and suppose that $(x_k, (g_k^i)_{i=1}^n)$ evolves according to Algorithm 1 with a constant lower bound $\varepsilon_k = \varepsilon \leq \frac{1}{2n}$ and a constant step size $\alpha_k = \alpha \leq \frac{1}{20}\frac{\varepsilon}{L}$. Then for any $k \in \mathbb{N}$ we have:*

$$\mathbb{E}\left[T^k\right] \leq (1 - \alpha\rho)^k T^0 + (1 + 2n\varepsilon)\frac{2\alpha\sigma_*^2}{\rho}$$

*Proof.* Applying Theorem 1 on $T^k$ repeatedly and bounding the resulting geometric sum we get the result. □