ECE 273 Convex Optimization and Applications — Instructor: Jun-Kun Wang
Scribe: Shaoxiu Wei, Jiajie Yu — April 25, 2024
Editor/TA: Marialena Sfyraki

Lecture 8: SGD with Variance Reduction

1 Review of Lecture 7: Introduction to stochastic optimization

We begin by reviewing the Stochastic Gradient Descent (SGD).

1.1 Stochastic Optimization

Consider $\min_{x\in\mathbb{R}^d} F(x)$, where $F(x) := \mathbb{E}_z[f(x;z)]$.

Algorithm: Stochastic Gradient Descent (SGD)

---

for $k = 1, 2, \ldots$ **do**
  Compute a stochastic gradient g_k that satisfies $\mathbb{E}_z[g_k] = \nabla F(x_k)$
  $x_{k+1} = x_k - \eta g_k$.
end for

1.2 SGD v.s. GD

	Method	
	SGD	**GD**
convex (and smooth)	$O\left(\frac{1}{\sqrt{k}}\right)$	$O\left(\frac{1}{k}\right)$
strongly convex and smooth	$O\left(\frac{1}{k}\right)$	$O\left(\exp(-k)\right)$

Table 1: Convergence rates for different optimization scenarios. [Rakhlin (2012)]

1.3 Iteration complexity of SGD

Denote $i_{1:K}$ all the randomness from iteration 1 to K, i.e., $i_1, i_2, \ldots, i_K$.

Theorem 1. *Let* $F(x) = \mathbb{E}_z[f(x;z)] : \mathbb{R}^d \to \mathbb{R}$ *be a convex function. Consider the update*

$$x_{k+1} = x_k - \eta g_k,$$

where $\mathbb{E}_z[g_k] = \nabla F(x_k)$. *Suppose* $x_* = \arg\min F(x)$ *exists and the initial distance is bounded, i.e.,* $\|x_1 - x_*\|_2 \le D$. *Then,*

$$\frac{1}{K}\sum_{k=1}^{K} \mathbb{E}_{i_{1:K}}\left[F(x_k) - F(x_*)\right] \le \frac{\eta}{2K}\left(\sum_{k=1}^{K} \mathbb{E}_{i_{1:K}}\left[\|g_k\|_2^2\right]\right) + \frac{\|x_1 - x_*\|_2^2}{2\eta K},$$

where $\bar{x}_K := \frac{1}{K}\sum_{k=1}^{K} x_k$.

Caution!!! To be rigorous, we need to identify conditions such that the stochastic gradient norm is bounded.

Lemma 1. $\mathbb{E}_{i_k}\left[\|g_k\|_2^2\right]$ *is an upper bound of the variance of the stochastic gradient.*

Proof.

$$\begin{aligned} Var(g_k) &\triangleq \mathbb{E}_{i_k}\left[(g_k - \mathbb{E}_{i_k}[g_k])^2\right] \\ &\big(\text{since } \mathbb{E}_{i_k}[g_k] = \nabla F(x_k)\big) \\ &= \mathbb{E}_{i_k}\left[\|g_k\|_2^2\right] - 2\mathbb{E}_{i_k}\left[\langle g_k, \mathbb{E}_{i_k}[g_k]\rangle\right] + \|\nabla F(x_k)\|_2^2 \\ &= \mathbb{E}_{i_k}\left[\|g_k\|_2^2\right] - 2\|\nabla F(x_k)\|_2^2 + \|\nabla F(x_k)\|_2^2 \\ &\le \mathbb{E}_{i_k}\left[\|g_k\|_2^2\right]. \end{aligned}$$

□

1.4 SGD for non-convex problems

Theorem 2. *Assume that the variance of the stochastic gradient* $\nabla f(x;z)$ *is at most* σ^2 *for all* x, *i.e.,* $\mathbb{E}_z\left[\|\nabla f(x;z) - \nabla F(x)\|_2^2\right] \le \sigma^2$. *Suppose* $F(\cdot)$ *is L-smooth. Then, SGD with the step size* $\eta \le \frac{1}{L}$ *has*

$$\sum_{k=1}^{K} \mathbb{E}_{i_{1:K}}\left[\|\nabla F(x_k)\|_2^2\right] \le \frac{2(F(x_1) - F_*)}{\eta} + \eta L \sigma^2 K.$$

Proof. (Proof of the theorem) Starting from the smoothness, we have, given x_k,

$$F(x_{k+1}) \le F(x_k) + \langle \nabla F(x_k), x_{k+1} - x_k\rangle + \frac{L}{2}\|x_{k+1} - x_k\|^2 \quad (1)$$

$$= F(x_k) - \eta\langle \nabla F(x_k), g_k\rangle + \frac{\eta^2 L}{2}\|g_k\|^2 \quad (2)$$

Take expectation over the randomness from 1 to k on both sides, we have

$$\mathbb{E}_{i_{1:k}}[F(x_{k+1})] \leq \mathbb{E}_{i_{1:k}}[F(x_k)] - \eta \mathbb{E}_{i_{1:k}}[\langle \nabla F(x_k), g_k \rangle] + \frac{\eta^2 L}{2} \mathbb{E}_{i_{1:k}}[\|g_k\|^2] \tag{3}$$

$$= \mathbb{E}_{i_{1:k}}[F(x_k)] - \eta \mathbb{E}_{i_{1:k-1}}[\|\nabla F(x_k)\|^2] + \frac{\eta^2 L}{2} \mathbb{E}_{i_{1:k}}[\|g_k\|^2] \tag{4}$$

$$\leq \mathbb{E}_{i_{1:k}}[F(x_k)] - \eta \mathbb{E}_{i_{1:k-1}}[\|\nabla F(x_k)\|^2] + \frac{\eta^2 L}{2} \left(\mathbb{E}_{i_{1:k-1}}[\|\nabla F(x_k)\|^2] + \sigma^2 \right) \tag{5}$$

$$\leq \mathbb{E}_{i_{1:k}}[F(x_k)] - \frac{\eta}{2} \mathbb{E}_{i_{1:k-1}}[\|\nabla F(x_k)\|^2] + \frac{\eta^2 L}{2} \sigma^2. \tag{6}$$

$$\leq \mathbb{E}_{i_{1:k}}[F(x_k)] - \frac{\eta}{2} \mathbb{E}_{i_{1:k}}[\|\nabla F(x_k)\|^2] + \frac{\eta^2 L}{2} \sigma^2. \tag{7}$$

It is noted that for (5), we used

$$\mathbb{E}_{i_{1:k}}[\|g_k\|^2] = \mathbb{E}_{i_{1:k-1}} \left[\mathbb{E}_{i_k}[\|g_k\|^2 | i_{1:k-1}] \right] \tag{8}$$

$$= \mathbb{E}_{i_{1:k-1}} \left[\mathbb{E}_{i_k}[\|g_k\|^2 | x_k] \right] \tag{9}$$

$$\leq \mathbb{E}_{i_{1:k-1}}[\|\nabla F(x_k)\|^2] + \sigma^2 \ ], \tag{10}$$

where the last inequality is by the assumption that the variance is bounded by σ^2. For (6), we used $\eta \leq \frac{1}{L}$. For (7), we used that i_k is independent from x_k.

Now take expectation over all the randomness on both sides of (7) and sum over $k = 1$ to K,

$$\mathbb{E}_{i_{1:K}} \left[F(x_{k+1}) - F(x_1) \right] \leq - \sum_{k=1}^{K} \frac{\eta}{2} \mathbb{E}_{i_{1:K}}[\|\nabla F(x_k)\|^2] + \frac{L\eta^2\sigma^2}{2}.$$

□

Corollary 2.1. *If $\hat{x}$ is selected uniformly at random from $x_1, \ldots, x_K$, then we have*

$$\mathbb{E}_{i_{1:K}} \left[\|\nabla F(\hat{x})\| \right] \leq \frac{\sqrt{2 \left(F(x_1) - F_* \right) L}}{\sqrt{K}} + \frac{\sqrt{3\sigma \sqrt{\left(F(x_1) - F_* \right) L}}}{K^{1/4}}. \tag{11}$$

2 SGD with variance reduction (SVRG)

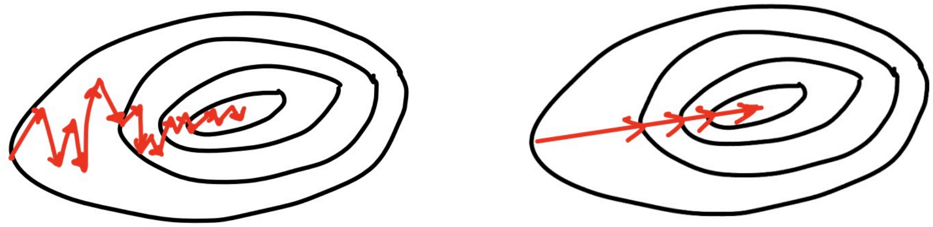

Figure 1: Progress of SGD (left) and GD (right) in practice

The variance of the stochastic gradient can be large. Thus, the question is **how to reduce the variance?**

2.1 SGD with variance reduction (SVRG) Algorithm

$$\min_{x \in \mathbb{R}^d} F(x), \text{ where } F(x) := \frac{1}{n}\sum_{i=1}^{n} f_i(x)$$

Algorithm: SGD with variance reduction (SVRG)

Set $s = 1$. Init $v_1 = x_1$. Learning rate η.
for stage $s = 1, 2, \ldots, S$ **do**
  for iteration $k = 1, 2, \ldots, K$ **do**
    Randomly pick a sample $i_k \in [n]$.
    Set $g_k = \nabla f_{i_k}(x_k) - \nabla f_{i_k}(v_s) + \nabla F(v_s)$. (variance reduction)
    Update $x_{k+1} = x_k - \eta g_k$.
  end for
  Update the snapshot $v_{s+1} = \frac{1}{k}\sum_{k=1}^{K} x_k$.
  Set $x_1 = v_{s+1}$
end for

2.2 Valid Stochastic Gradient

Let's show that it is a valid stochastic gradient.

Lemma 2. (**Unbiased Estimate**)

$$\mathbb{E}_{i_k}\left[\nabla f_{i_k}(x_k) - \nabla f_{i_k}(v_s) + \nabla F(v_s)\right] = \nabla F(x_k). \tag{12}$$

Proof.

$$\mathbb{E}_{i_k}\left[\nabla f_{i_k}(x_k) - \nabla f_{i_k}(v_s) + \nabla F(v_s)\right] = \nabla F(x_k) - \nabla F(v_s) + \nabla F(v_s) = \nabla F(x_k).$$

□

Recall that $\mathbb{E}_{i_k}\left[\|g_k\|_2^2\right]$ is an upper bound of the variance of $g_k \in \mathbb{R}^d$. Let us analyze the squared gradient norm $\mathbb{E}_{i_k}[\|g_k\|_2^2]$.

Lemma 3. (**Variance bound**)

$$\mathbb{E}_{i_k}\left[\|g_k\|_2^2\right] \leq 4L\left(F(x_k) - F(x_*)\right) + 4L\left(F(v_s) - F(x_*)\right). \tag{13}$$

Before we begin the proof of Lemma 3, we need to introduce two additional lemmas as below:

Lemma 4. *For any random variable $Y \in \mathbb{R}^d$,*

$$\mathbb{E}\left[\|Y - \mathbb{E}[Y]\|_2^2\right] = \mathbb{E}[\|Y\|_2^2] - \left(\mathbb{E}[\|Y\|]\right)^2 \leq \mathbb{E}[\|Y\|_2^2]. \tag{14}$$

Lemma 5. *If each $f_i(\cdot)$ is L-smooth convex, then*

$$\mathbb{E}_{i_k}\left[\|\nabla f_{i_k}(x) - \nabla f_{i_k}(x_*)\|^2\right] \leq 2L\left(F(x) - F(x_*)\right). \tag{15}$$

Proof. We will proof Lemma 5 in Homework 3. □

Proof. (Proof of Lemma 3)

$$\begin{aligned}
\mathbb{E}_{i_k}\left[\|g_k\|_2^2\right] &= \mathbb{E}_{i_k}\left[\|\nabla f_{i_k}(x_k) - \nabla f_{i_k}(v_s) + \nabla F(v_s)\|_2^2\right] \\
&= \mathbb{E}_{i_k}\left[\|\nabla f_{i_k}(x_k) - \nabla f_{i_k}(x_*) + \nabla f_{i_k}(x_*) - \nabla f_{i_k}(v_s) + \nabla F(v_s)\|_2^2\right] \\
&\leq 2\mathbb{E}_{i_k}\left[\|\nabla f_{i_k}(x_k) - \nabla f_{i_k}(x_*)\|_2^2\right] \\
&\quad + 2\mathbb{E}_{i_k}\left[\|\nabla f_{i_k}(x_*) - \nabla f_{i_k}(v_s) + \nabla F(v_s)\|_2^2\right],
\end{aligned}$$

where the last inequality follows from $||a+b||_2^2 \leq 2||a||_2^2 + 2||b||_2^2$. Based on Lemma 4, we can further rewrite the second term in above inequality as

$$\begin{aligned}
&2\mathbb{E}_{i_k}\left[\|\nabla f_{i_k}(x_*) - \nabla f_{i_k}(v_s) + \nabla F(v_s)\|_2^2\right] \\
&= 2\mathbb{E}_{i_k}\left[\|\underbrace{\nabla f_{i_k}(x_*) - \nabla f_{i_k}(v_s)}_{:=Y} - \underbrace{\left(\nabla F(x_*) - \nabla F(v_s)\right)}_{:=E[Y]}\|_2^2\right] \\
&\leq 2\mathbb{E}_{i_k}\left[\underbrace{\|\nabla f_{i_k}(x_*) - \nabla f_{i_k}(v_s)\|_2^2}_{:=||Y||_2^2}\right].
\end{aligned}$$

Therefore,

$$\mathbb{E}_{i_k}\left[\|g_k\|_2^2\right] \le 2\mathbb{E}_{i_k}\left[\|\nabla f_{i_k}(x_k) - \nabla f_{i_k}(x_*)\|_2^2\right] + 2\mathbb{E}_{i_k}\left[\|\nabla f_{i_k}(v_s) - \nabla f_{i_k}(x_*)\|_2^2\right].$$

By using Lemma 5, above equation could be further lead to

$$\mathbb{E}_{i_k}[\|g_k\|^2] \le 4L\left(F(x_k) - F(x_*)\right) + 4L\left(F(v_s) - F(x_*)\right).$$

□

2.3 Convergence for each stage

Recall $s \in [S]$ is the index of a stage. Denote z_s all the randomness (in the inner iterations) at stage s.

Theorem 3. *Suppose each $f_i(\cdot)$ is L-smooth and μ-strongly convex. Setting $\eta = \frac{1}{8L}$ and $K = 64\frac{L}{\mu}$. Then, at each stage s,*

$$\mathbb{E}_{z_s}\left[F(v_{s+1}) - F(x_*)\right] \le \frac{3}{4}\left(F(v_s) - F(x_*)\right), \tag{16}$$

where $x_* \in \arg\min_x F(x)$.

Proof. By Theorem 1, we have

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}_{z_s}\left[F(x_k) - F(x_*)\right] \le \frac{\eta}{2K}\left(\sum_{k=1}^{K}\mathbb{E}_{z_s}\left[\|g_k\|_2^2\right]\right) + \frac{F(x_1) - F(x_*)}{\eta\mu K}. \tag{17}$$

Recall the lemma of the variance bound, i.e., Lemma 3, we have, given x_k and v_s,

$$\mathbb{E}_{i_k}[\|g_k\|_2^2] \le 4L\left(F(x_k) - F(x_*)\right) + 4L\left(F(v_s) - F(x_*)\right). \tag{18}$$

Taking expectation over all the randomness at stage s on both sides of (18) further, we have

$$\mathbb{E}_{z_s}[\|g_k\|_2^2] \le \mathbb{E}_{z_s}\left[4L\left(F(x_k) - F(x_*)\right) + 4L\left(F(v_s) - F(x_*)\right)\right]. \tag{19}$$

Summing (19) over all the inner iterations at stage s, we get

$$\frac{\eta}{2K}\sum_{k=1}^{K}\mathbb{E}_{z_s}[\|g_k\|^2] \le \frac{\eta}{2K}\sum_{k=1}^{K}\mathbb{E}_{z_s}\left[4L\left(F(x_k) - F(x_*)\right) + 4L\left(F(x_1) - F(x_*)\right]\right). \tag{20}$$

Combining (17) and (20), we have

$$(1-2\eta L)\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}_{z_s}\left[F(x_k)-F(x_*)\right] \le \left(2\eta L+\frac{1}{\eta\mu K}\right)\big(F(x_1)-F(x_*)\big). \tag{21}$$

Setting $\eta = \frac{1}{8L}$ and $K = 64\frac{L}{\mu}$, we have

$$\mathbb{E}_{z_s}\left[F(\bar{x}_K)-F(x_*)\right] \le \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}_{z_s}\left[F(x_k)-F(x_*)\right] \le \frac{3}{4}\big(F(x_1)-F(x_*)\big), \tag{22}$$

where the first inequality follows from Jensen's inequality.

By the algorithm design, we have that x_1 is equal to the snapshot v_s at each stage s. Therefore, $\frac{3}{4}\big(F(x_1)-F(x_*)\big) = \frac{3}{4}\big(F(v_s)-F(x_*)\big)$. Additionally, $\bar{x}_k$ is used to initialize x_1 and the snapshot v_{s+1} in the next stage. Thus, $\mathbb{E}_{z_s}\left[F(\bar{x}_K)-F(x_*)\right] = \mathbb{E}_{z_s}\left[F(v_{s+1})-F(x_*)\right]$. We can hence re-write (22) as

$$\mathbb{E}_{z_s}\left[F(v_{s+1})-F(x_*)\right] \le \frac{3}{4}\big(F(v_s)-F(x_*)\big).$$

Thus, the above means that the expected gap is shrinking within a constant factor in each stage.

□

3 Complexity Analysis

3.1 Iteration complexity of SVRG

To get an expected ϵ-gap, the total number of stages is:

$$\begin{aligned}
&\mathbb{E}_{z_{1:s}}\left[F(v_{s+1})-F(x_*)\right] \\
&= \sum_{\cdot}\Pr\left(z_{1:s-1}=\cdot\right)\mathbb{E}_{z_s}\left[F(v_{s+1})-F(x_*)|z_{1:s-1}=\cdot\right] \qquad (23)\\
&= \mathbb{E}_{z_{1:s-1}}\left[\ \mathbb{E}_{z_s}\left[F(v_{s+1})-F(x_*)|z_{1:s-1}\right]\ \right].
\end{aligned}$$

According to Theorem 3, we have

$$\begin{aligned}
\mathbb{E}_{z_{1:s}}\left[F(v_{s+1})-F(x_*)\right] &\le \frac{3}{4}\mathbb{E}_{z_{1:s}}\left[F(v_s)-F(x_*)\right] \\
&\le \left(\frac{3}{4}\right)^S\big(F(v_1)-F(x_*)\big) \\
&\le \epsilon
\end{aligned}$$

$$\Leftrightarrow S \geq 4 \log \left(\frac{F(v_1) - F(x_*)}{\epsilon} \right) = \mathcal{O} \left(\log \left(\frac{1}{\epsilon} \right) \right)$$

. According to the above calculations, the total number of stochastic gradient computations could be represented by

$$2 \times K \times S = \mathcal{O} \left(\frac{L}{\mu} \log \left(\frac{1}{\epsilon} \right) \right). \tag{24}$$

Total number of Full gradient computations is

$$S = \mathcal{O} \left(n \log \left(\frac{1}{\epsilon} \right) \right). \tag{25}$$

Here, the cost of the full gradient computation = cost of n stochastic gradient computation. Total number of (equivalent) stochastic gradient computations is

$$\mathcal{O} \left(\frac{L}{\mu} \log \left(\frac{1}{\epsilon} \right) \right) + \mathcal{O} \left(n \log \left(\frac{1}{\epsilon} \right) \right). \tag{26}$$

3.2 SVRG v.s. GD

Therefore, we could obtain

$$\begin{aligned} &\frac{\text{runtime of SVRG}}{\text{runtime of GD}} \\ &= \frac{\#\text{ (equivalent) stochastic gradient computs. of SVRG}}{\#\text{ (equivalent) stochastic gradient computs. of GD}} \\ &= \frac{\left(\frac{L}{\mu} + n \right) \log(\frac{1}{\epsilon})}{\frac{L}{\mu} \log\left(\frac{1}{\epsilon}\right) \times n}. \end{aligned} \tag{27}$$

Lets prove that the runtime of SVRG is generally smaller than runtime of GD, i.e.

$$\begin{aligned} \frac{\text{runtime of SVRG}}{\text{runtime of GD}} \leq 1 \Leftrightarrow \frac{\left(\frac{L}{\mu} + n \right) \log\left(\frac{1}{\epsilon}\right)}{\frac{L}{\mu} \log\left(\frac{1}{\epsilon}\right) \times n} \leq 1 \\ \Leftrightarrow n \leq (n-1) \frac{L}{\mu} \\ \Leftrightarrow \mu \leq L, \text{ as } n \to \infty \end{aligned}$$

which is always true. The proof is completed.

3.3 SVRG v.s. SGD

$$\frac{\text{runtime of SVRG}}{\text{runtime of SGD}} = \frac{\left(\frac{L}{\mu} + n\right)\log\left(\frac{1}{\epsilon}\right)}{\frac{1}{\epsilon} \times 1}.$$

The condition for SVRG to be faster than SGD than SGD is when

$$\left(\frac{L}{\mu} + n\right)\log\left(\frac{1}{\epsilon}\right) \ll \frac{1}{\epsilon} \Leftrightarrow \left(\frac{L}{\mu} + n\right) \ll \frac{\frac{1}{\epsilon}}{\log\frac{1}{\epsilon}}.$$

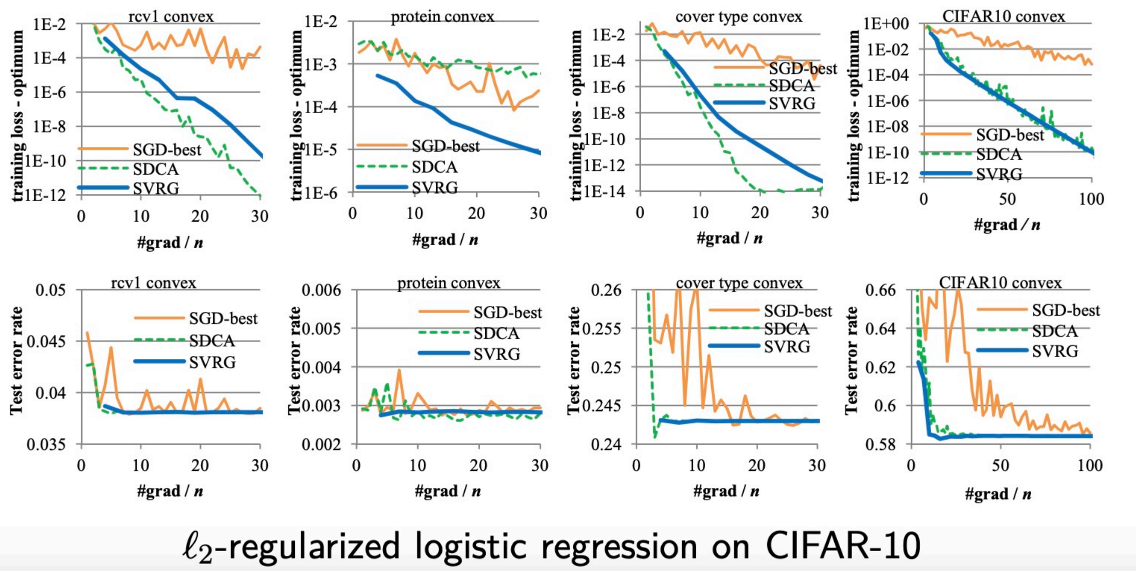

Figure 2: ℓ_2 - regularized logistic regression on CIFAR-10. [Johnson (2013)]

Bibliographic notes

More information can be found in [Drusvyatskiy (2020)], [Vishnoi (2021)], [Rakhlin (2012)], and [Johnson (2013)].

References

[Drusvyatskiy (2020)] Dmitriy Drusvyatskiy. Convex Analysis and Nonsmooth Optimization. 2020.

[Vishnoi (2021)] Nisheeth K. Vishnoi. Algorithms for Convex Optimization. Cambridge University Press, 2021.

[Rakhlin (2012)] Alexander Rakhlin. Making Gradient Descent Optimal for Strongly Convex Stochastic Optimization. ICML, 2012.

[Johnson (2013)] Rie Johnson. Accelerating Stochastic Gradient Descent using Predictive Variance Reduction. NeurIPS, 2013.