Stop Wasting My Gradients: Practical SVRG Appendix

Reza Babanezhad, Mohamed Osama Ahmed, Alim Virani, Mark Schmidt
Department of Computer Science
University of British Columbia

Jakub Konečný
School of Mathematics
University of Edinburgh

Scott Sallinen
Department of Electrical and Computer Engineering
University of British Columbia

A Convergence Rate of SVRG with Error

We first give the proof of Proposition 1, which gives a convergence rate for SVRG with an error and uniform sampling. We then turn to the case of non-uniform sampling.

A.1 Proof of Proposition 1

We follow a similar argument to Johnson & Zhang [8], but propagating the error e^s through the analysis. We begin by deriving a simple bound on the variance of the sub-optimality of the gradients.

Lemma 1. *For any x,*

$$\frac{1}{n}\sum_{i=1}^{n}\|f_i'(x) - f_i'(x^*)\|^2 \leq 2L[f(x) - f(x^*)].$$

Proof. Because each f_i' is L-Lipschitz continuous, we have [11, Theorem 2.1.5]

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

Setting $y = x^*$ and summing this inequality times $(1/n)$ over all i we obtain the result. □

In this section we'll use $\tilde{x}$ to denote x^s, e to denote e^s, and we'll use ν_t to denote the search direction at iteration t,

$$\nu_t = f_{i_t}'(x_{t-1}) - f_{i_t}'(\tilde{x}) + \tilde{\mu} + e.$$

Note that $\mathbb{E}[\nu_t] = f'(x_{t-1}) + e$ and the next lemma bounds the variance of this value.

Lemma 2. *In each iteration t of the inner loop,*

$$\mathbb{E}\|\nu_t\|^2 \leq 4L[f(x_{t-1}) - f(x^*)] + 4L[f(\tilde{x}) - f(x^*)] + 2\|e\|^2$$

Proof. By using the inequality $\|x + y\|^2 \leq 2\|x\|^2 + 2\|y\|^2$ and the property

$$\mathbb{E}[f_{i_t}'(\tilde{x}) - f_{i_t}'(x^*)] = f'(\tilde{x}),$$

we have

$$\begin{aligned}
\mathbb{E}\|\nu_t\|^2 &= \mathbb{E}\|f'_{i_t}(x_{t-1}) - f'_{i_t}(\tilde{x}) + \tilde{\mu} + e\|^2 \\
&\leq 2\mathbb{E}\|f'_{i_t}(x_{t-1}) - f'_{i_t}(x^*)\|^2 + 2\mathbb{E}\|[f'_{i_t}(\tilde{x}) - f'_{i_t}(x^*)] - f'(\tilde{x}) - e\|^2 \\
&= 2\mathbb{E}\|f'_{i_t}(x_{t-1}) - f'_{i_t}(x^*)\|^2 + 2\mathbb{E}\|[f'_{i_t}(\tilde{x}) - f'_{i_t}(x^*)] - \mathbb{E}[f'_{i_t}(\tilde{x}) - f'_{i_t}(x^*)] - e\|^2 \\
&= 2\mathbb{E}\|f'_{i_t}(x_{t-1}) - f'_{i_t}(x^*)\|^2 + 2\mathbb{E}\|[f'_{i_t}(\tilde{x}) - f'_{i_t}(x^*)] - \mathbb{E}[f'_{i_t}(\tilde{x}) - f'_{i_t}(x^*)]\|^2 + 2\|e\|^2 \\
&\quad - 4\mathbb{E}\left\langle [f'_{i_t}(\tilde{x}) - f'_{i_t}(x^*)] - \mathbb{E}[f'_{i_t}(\tilde{x}) - f'_{i_t}(x^*)], e\right\rangle \\
&= 2\mathbb{E}\|f'_{i_t}(x_{t-1}) - f'_{i_t}(x^*)\|^2 + 2\mathbb{E}\|[f'_{i_t}(\tilde{x}) - f'_{i_t}(x^*)] - \mathbb{E}[f'_{i_t}(\tilde{x}) - f'_{i_t}(x^*)]\|^2 + 2\|e\|^2,
\end{aligned}$$

If we now use that $\mathbb{E}[\|X - \mathbb{E}[X]\|^2] \leq \mathbb{E}\|X\|^2$ for any random variable X, we obtain the result by applying Lemma 1 to bound $\mathbb{E}\|f'_{i_t}(x_{t-1}) - f'_{i_t}(x^*)\|^2$ and $\mathbb{E}\|[f'_{i_t}(\tilde{x}) - f'_{i_t}(x^*)]\|^2$. □

The following Lemma gives a bound on the distance to the optimal solution.

Lemma 3. *In every iteration t of the inner loop,*

$$\begin{aligned}
\mathbb{E}\|x_t - x^*\|^2 \leq \|x_{t-1} - x^*\|^2 - 2\eta(1 - 2\eta L)[f(x_{t-1}) - f(x^*)] \\
+ 4L\eta^2[f(\tilde{x}) - f(x^*)] + 2\eta(Z\|e\| + \eta\|e\|^2).
\end{aligned}$$

Proof. We expand the expectation and bound $\mathbb{E}\|\nu_t\|^2$ using Lemma 2 to obtain

$$\begin{aligned}
&\mathbb{E}\|x_t - x^*\|^2 \\
&\quad = \|x_{t-1} - x^*\|^2 - 2\eta\left\langle x_{t-1} - x^*, \mathbb{E}[\nu_t]\right\rangle + \eta^2\mathbb{E}\|\nu_t\|^2 \\
&\quad = \|x_{t-1} - x^*\|^2 - 2\eta\left\langle x_{t-1} - x^*, f'(x_{t-1}) + e\right\rangle + \eta^2\mathbb{E}\|\nu_t\|^2 \\
&\quad = \|x_{t-1} - x^*\|^2 - 2\eta\left\langle x_{t-1} - x^*, f'(x_{t-1})\right\rangle - 2\eta\left\langle x_{t-1} - x^*, e\right\rangle + \eta^2\mathbb{E}\|\nu_t\|^2 \\
&\quad \leq \|x_{t-1} - x^*\|^2 - 2\eta\left\langle x_{t-1} - x^*, e\right\rangle - 2\eta[f(x_{t-1}) - f(x^*)] + 2\eta^2\|e\|^2 \\
&\qquad + 4L\eta^2[f(x_{t-1}) - f(x^*)] + 4L\eta^2[f(\tilde{x}) - f(x^*)]
\end{aligned}$$

The inequality above follows from convexity of f. The result follows from applying Cauchy-Schwartz to the linear term in e and that $\|x_{t-1} - x^*\| \leq Z$. □

To prove Proposition 1 from the main paper, we first sum the inequality in Lemma 3 for all $t = 1, ..., m$ and take the expectation with respect to the choice of x^s to get

$$\begin{aligned}
\mathbb{E}\|x_m - x^*\|^2 \leq \mathbb{E}\|x_0 - x^*\|^2 - 2\eta(1 - 2L\eta)m\mathbb{E}[f(x_{t-1}) - f(x^*)] \\
+ 4L\eta^2 m\mathbb{E}[f(\tilde{x}) - f(x^*)] + 2m\eta(Z\mathbb{E}\|e\| + \eta\mathbb{E}\|e\|^2).
\end{aligned}$$

Re-arranging, and noting that $x_0 = \tilde{x}_{s-1}$ and $\mathbb{E}\left[f(x_{t-1})\right] = \mathbb{E}\left[f(x^s)\right]$, we have that

$$\begin{aligned}
&2\eta(1 - 2L\eta)m\mathbb{E}[f(x_s) - f(x^*)] \\
&\qquad \leq \mathbb{E}\|\tilde{x}_{s-1} - x^*\|^2 + 4L\eta^2 m\mathbb{E}[f(\tilde{x}_{s-1}) - f(x^*)] + 2m\eta(Z\mathbb{E}\|e^{s-1}\| + \eta\mathbb{E}\|e^{s-1}\|^2) \\
&\qquad \leq \frac{2}{\mu}\mathbb{E}[f(\tilde{x}_{s-1}) - f(x^*)] + 4L\eta^2 m\mathbb{E}[f(\tilde{x}_{s-1}) - f(x^*)] + 2m\eta(Z\mathbb{E}\|e\| + \eta\mathbb{E}\|e\|^2),
\end{aligned}$$

where the last inequality uses strong-convexity and that $f'(x^*) = 0$. By dividing both sides by $2\eta(1 - 2L\eta)m$ (which is positive due to the constraint $\eta \leq 1/2L$ implied by $0 < \rho < 1$ and $\eta > 0$), we get

$$\begin{aligned}
&\mathbb{E}[f(x_s) - f(x^*)] \\
&\leq \left(\frac{1}{m\mu(1 - 2\eta L)\eta} + \frac{2L\eta}{1 - 2\eta L}\right)\mathbb{E}[f(\tilde{x}_{s-1}) - f(x^*)] + \frac{1}{1 - 2\eta L}\left(Z\mathbb{E}\|e^{s-1}\| + \eta\mathbb{E}\|e^{s-1}\|^2\right).
\end{aligned}$$

A.2 Non-Uniform Sampling

If we sample i_t proportional to the individual Lipschitz constants L_i, then we have the following analogue of Lemma 1.

Lemma 4. *For any x,*

$$\mathbb{E}\left\|\frac{L_i}{\bar{L}}[f_i'(x) - f_i'(x^*)]\right\|^2 \leq 2\bar{L}[f(x) - f(x^*)].$$

Proof. Because each f_i' is L_i-Lipschitz continuous, we have [11, Theorem 2.1.5]

$$f_i(x) \geq f_i(y) + \langle f_i'(x), x - y\rangle + \frac{1}{2L_i}\|f_i'(x) - f_i'(y)\|^2.$$

Setting $y = x^*$ and summing this inequality times $(1/n)$ over all i we have

$$\begin{aligned}
&\mathbb{E}\left\|\frac{L_i}{\bar{L}}[f_i'(x) - f_i'(x^*)]\right\|^2 = \sum_{i=1}^n \frac{L_i}{n\bar{L}}\frac{\bar{L}^2}{L_i^2}\|f_i'(x) - f_i'(y)\|^2 = \frac{\bar{L}}{n}\sum_{i=1}^n \frac{1}{L_i}\|f_i'(x) - f_i'(y)\|^2 \\
&\leq \frac{\bar{L}}{n}\sum_{i=1}^n \frac{1}{L_i} 2L_i[f_i(x) - f_i(x^*) - \langle f_i'(x), x - x^*\rangle] \\
&= 2\bar{L}[f(x) - f(x^*)]
\end{aligned}$$

□

With this modified lemma, we can derive the convergence rate under this non-uniform sampling scheme by following an identical sequence of steps but where each instance of L is replaced by $\bar{L}$.

B Mixed SVRG and SG Method

We first give the proof of Proposition 2 in the paper, which analyzes a method that mixes SG and SVRG updates using a constant step size. We then consider a variant where the SG and SVRG updates use different step sizes.

B.1 Proof of Proposition 2

Recall that the SG update is

$$x_t = x_{t-1} - \eta f_{i_t}'(x_{t-1}).$$

Using this in Lemma 3 and following a similar argument we have

$$\begin{aligned}
\mathbb{E}\|x_t - x^*\|^2 &\leq \alpha\{\|x_{t-1} - x^*\|^2 - 2\eta(1 - 2\eta L)[f(x_{t-1}) - f(x^*)] + 4L\eta^2[f(\tilde{x}) - f(x^*)] + 2\eta(Z\|e\| + \eta\|e\|^2)\} \\
&\quad + \beta\{\|x_{t-1} - x^*\|^2 + \eta^2\mathbb{E}\|f_{i_t}'(x_{t-1})\|^2 - 2\eta\left\langle x_{t-1} - x^*, \mathbb{E}[f_{i_t}'(x_{t-1})]\right\rangle\} \\
&\leq \|x_{t-1} - x^*\|^2 - 2\eta(1 - 2\eta L)[f(x_{t-1}) - f(x^*)] + \alpha 4L\eta^2[f(\tilde{x}) - f(x^*)] + \alpha 2\eta(Z\|e\| + \eta\|e\|^2) + \beta\eta^2\sigma^2,
\end{aligned}$$

where the second inequality uses convexity of f and we have defined $\beta = (1 - \alpha)$. We now sum up both sides and take the expectation with respect to the history,

$$\begin{aligned}
\mathbb{E}\|x_m - x^*\|^2 \leq \mathbb{E}\|x_0 - x^*\|^2 &- 2\eta(1 - 2L\eta)m\mathbb{E}[f(x_{t-1}) - f(x^*)] \\
&+ 4\alpha L\eta^2 m\mathbb{E}[f(\tilde{x}) - f(x^*)] + 2m\alpha\eta(Z\mathbb{E}\|e\| + \eta\mathbb{E}\|e\|^2) \\
&+ m\beta\eta^2\sigma^2.
\end{aligned}$$

By re-arranging the terms we get

$$\begin{aligned}
2\eta(1 - 2L\eta)m\mathbb{E}[f(x_s) - f(x^*)] \leq \frac{2}{\mu}\mathbb{E}[f(\tilde{x}_{s-1}) - f(x^*)] &+ 4\alpha L\eta^2 m\mathbb{E}[f(\tilde{x}_{s-1}) - f(x^*)] \\
&+ 2m\alpha\eta(Z\mathbb{E}\|e\| + \eta\mathbb{E}\|e\|^2) + m\beta\eta^2\sigma^2,
\end{aligned}$$

and by dividing both sides by $2\eta(1 - 2L\eta)m$ we get the result.

B.2 Mixed SVRG and SG with Different Step Sizes

Consider a variant where we use a step size of η in the SVRG update and a step-size η_s in the SG update (which will decrease as the iterations proceed). Analyzing the mixed algorithm in this setting gives

$$\begin{aligned}
\mathbb{E}\|x_t - x^*\|^2 &\leq \alpha\{\|x_{t-1} - x^*\|^2 - 2\eta(1-2\eta L)[f(x_{t-1}) - f(x^*)] \\
&\quad + 4L\eta^2[f(x^s) - f(x^*)] + 2\eta(Z\|e^s\| + \eta\|e^s\|^2)\} \\
&\quad + \beta\{\|x_{t-1} - x^*\|^2 + \eta_s^2\mathbb{E}\|f'_{i_t}(x_{t-1})\|^2 - 2\eta_s\left\langle x_{t-1} - x^*, \mathbb{E}[f'_{i_t}(x_{t-1})]\right\rangle\} \\
&= \mathbb{E}\left[\|x_{t-1} - x^*\|^2\right] - 2\alpha\eta(1-2\eta L)[f(x_{t-1}) - f(x^*)] \\
&\quad + 4\alpha L\eta^2[f(x^s) - f(x^*)] + 2\alpha\eta(Z\|e^s\| + \eta\|e^s\|^2) \\
&\quad + \beta\eta_s^2\mathbb{E}\|f'_{i_t}(x_{t-1})\|^2 - 2\beta\eta_s\left\langle x_{t-1} - x^*, f(x_{t-1})\right]\rangle \\
&\leq \mathbb{E}\left[\|x_{t-1} - x^*\|^2\right] - \{2\alpha\eta(1-2\eta L) + 2\beta\eta_s\}[f(x_{t-1}) - f(x^*)] \\
&\quad + 4\alpha L\eta^2[f(x^s) - f(x^*)] + 2\alpha\eta(Z\|e^s\| + \eta\|e^s\|^2) + \beta\eta_s^2\sigma^2.
\end{aligned}$$

As before, we take the expectation for all t and sum up these values, then rearranage and use strong-convexity of f to get

$$\begin{aligned}
&2m\{\alpha\eta(1-2\eta L) + \beta\eta_s\}[f(x_s) - f(x^*)] \\
&\quad \leq \left\{\frac{2}{\mu} + 4m\alpha L\eta^2\right\}[f(x^s) - f(x^*)] + 2m\alpha\eta(Z\|e^s\| + \eta\|e^s\|^2) + m\beta\eta_s^2\sigma^2.
\end{aligned}$$

If we now divide both side by $2m(\alpha\eta(1-2\eta L) + \beta\eta_s)$, we get

$$\begin{aligned}
&\mathbb{E}\left[f(x_s) - f(x^*)\right] \\
&\leq \Big\{\frac{1}{\mu m(\alpha\eta(1-2\eta L) + \beta\eta_s)} + \frac{2\alpha L\eta^2}{\alpha\eta(1-2\eta L) + \beta\eta_s}\Big\}[f(x^s) - f(x^*)] \\
&\quad + \frac{\alpha\eta}{\alpha\eta(1-2\eta L) + \beta\eta_s}(Z\mathbb{E}\left[\|e^s\|\right] + \eta\mathbb{E}\left[\|e^s\|^2\right]) + \frac{1}{2(\alpha\eta(1-2\eta L) + \beta\eta_s)}\beta\eta_s^2\sigma^2.
\end{aligned}$$

To improve the dependence on the error e^s and variance σ^2 compared to the basic SVRG algorithm with error e^s (Proposition 1), we require that the terms depending on these values are smaller,

$$\begin{aligned}
&\frac{\alpha\eta}{\alpha\eta(1-2\eta L) + \beta\eta_s}\left(Z\mathbb{E}\left[\|e^s\|\right] + \eta\mathbb{E}\left[\|e^s\|^2\right]\right) \\
&+ \frac{1}{2(\alpha\eta(1-2\eta L) + \beta\eta_s)}\beta\eta_s^2\sigma^2 \leq \frac{1}{1-2\eta L}\left(Z\mathbb{E}\left[\|e^s\|\right] + \eta\mathbb{E}\left[\|e^s\|^2\right]\right).
\end{aligned}$$

Let $\kappa = (1-2\eta L)$ and $\zeta = Z\mathbb{E}\left[\|e^s\|\right] + \eta\mathbb{E}\left[\|e^s\|^2\right]$, this requires

$$\frac{\alpha\eta}{\alpha\eta\kappa + \beta\eta_s}\zeta + \frac{\beta\eta_s^2}{2(\alpha\eta\kappa + \beta\eta_s)}\sigma^2 \leq \frac{\zeta}{\kappa}.$$

Thus, it is sufficient that η_s satisfies

$$\eta_s \leq \frac{2\zeta}{\kappa\sigma^2}.$$

Using the relationship between expected error and S^2, while noting that $S^2 \leq \sigma^2$ and $\frac{(n-|\mathcal{B}|)}{n|\mathcal{B}|} \leq 1$, a step size of the form $\eta_s = O^*(\sqrt{(n-|\mathcal{B}|)/n|\mathcal{B}|})$ will improve the dependence on e^s and σ^2 compared to the dependence on e^s in the pure SVRG method.

Algorithm 1 Batching Prox SVRG

Input: update frequency m and learning rate η and sample size increasing rate α
Initialize $\tilde{x}$
for $s = 1, 2, 3, \ldots$ **do**
  Choose batch size $|\mathcal{B}|$
  $\mathcal{B}$ = randomly choose $|\mathcal{B}|$ elements of $\{1, 2, \ldots, n\}$.
  $\tilde{\mu} = \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} g_i'(\tilde{x})$
  $x_0 = \tilde{x}$
  for $t = 1, 2, \ldots, m$ **do**
    Randomly pick $i_t \in 1, \ldots, n$
    $\nu_t = g_{i_t}'(x_{t-1}) - g_{i_t}'(\tilde{x}) + \tilde{\mu}$
    $x_t = prox_{\eta h}(x_{t-1} - \eta \nu_t)$ $(*)$
  end for
  set $\tilde{x} = \frac{1}{m} \sum_{t=1}^{m} x_t$
end for

C Proximal and Regularized SVRG

In this section we consider objectives of the form

$$f(x) = h(x) + g(x),$$

where $g(x) = \frac{1}{n} \sum_{i=1}^{n} g_i(x)$. We first consider the case where h is non-smooth and consider a proximal-gradient variant of SVRG where there is an error in the calculation of the gradient (Algorithm 1). We then consider smooth functions h where we use a modified SVRG iteration,

$$x_{t+1} = x_t - \eta \left(h'(x_t) + g_{i_t}'(x_t) - g_{i_t}'(x^s) + \mu^s \right),$$

where $\mu^s = g(x^s)$.

C.1 Composite Case

Similar to the work of [12], in this section we assume that f, g and h are μ-, μ_g-, μ_h-strongly convex (respectively). As before, we assume each g_i is convex and has an L-Lipschitz continuous gradient, but h can potentially be non-smooth. The algorithm we propose here extends the algorithm of [12], but adding an error term. In the algorithm we use the proximal operator which is defined by

$$\operatorname*{prox}_h(y) = \operatorname*{arg\,min}_{x \in \mathbb{R}^p} \{ \frac{1}{2} \|x - y\|^2 + h(x) \}$$

Below, we give a convergence rate for this algorithm with an error e^s.

Proposition 5. *If we have $\tilde{\mu}_s = g'(x^s) + e^s$ and set the step-size η and number of inner iterations m so that*

$$\rho \equiv \frac{1}{m\mu(1 - 4\eta L)\eta} + \frac{4L\eta(m+1)}{(1 - 4\eta L)m} < 1,$$

then Algorithm 1 has

$$\mathbb{E}[f(x_{s+1}) - f(x^*)] \leq \rho E[f(\tilde{x}_s) - f(x^*)] + \frac{1}{1 - 4\eta L} (Z\mathbb{E}\|e^s\| + \eta \mathbb{E}\|e^s\|^2,$$

where $\|x_t - x^*\| < Z$

To prove Proposition 5, we use Lemma 1,2 and 3 from [12], which are unchanged when we allow an error. Below we modify their Corollary 3 and then the proof of their main theorem.

Lemma 5. *Consider* $\nu_t = g_{i_t}^{'}(x_{t-1}) - g_{i_t}^{'}(\tilde{x}) + g'(\tilde{x}) + e$. *Then,*

$$\mathbb{E}\|\nu_t - g'(x_{t-1})\|^2 \leq \|e\|^2 + 4L[f(x_{t-1}) - f(x^*) + f(\tilde{x}) - f(x^*)]$$

Proof.

$$\begin{aligned}
&\mathbb{E}\|\nu_t - g'(x_{t-1})\|^2 \\
&= \mathbb{E}\|g_{i_t}^{'}(x_{t-1}) - g_{i_t}^{'}(\tilde{x}) + g'(\tilde{x}) + e - g'(x_{t-1})\|^2 \\
&= \|e\|^2 + E\|g_{i_t}^{'}(x_{t-1}) - g_{i_t}^{'}(\tilde{x}) + g'(\tilde{x}) - g'(x_{t-1})\|^2 \\
&\leq \|e\|^2 + E\|g_{i_t}^{'}(x_{t-1}) - g_{i_t}^{'}(\tilde{x})\|^2 \\
&\leq \|e\|^2 + 2E\|g_{i_t}^{'}(x_{t-1}) - g_{i_t}^{'}(x^*)\|^2 + 2E\|g_{i}^{'}(\tilde{x}) - g_{i_t}^{'}(x^*)\|^2
\end{aligned}$$

Using Lemma 1 from [12] and bounding the two expectations gives the result. □

Now we turn to proving Proposition 5.

Proof. Following the proof of Theroem 1 in [12], we have

$$\|x_t - x^*\|^2 \leq \|x_{t-1} - x^*\|^2 - 2\eta[f(x_t) - f(x^*)] - 2\eta\left\langle \Delta_t, x_t - x^* \right\rangle$$

where $\Delta_t = \nu_t - g'(x_{t-1})$ and $\mathbb{E}[\Delta_t] = e$. Now to bound $\langle \Delta_t, x_t - x^* \rangle$, we define

$$\bar{x_t} = prox_h(x_{t-1} - \eta g'(x_{t-1})),$$

and subsequently that

$$-2\eta\left\langle \Delta_t, x_t - x^* \right\rangle \leq 2\eta^2\|\Delta_t\|^2 - 2\eta\left\langle \Delta_t, \bar{x}_t - x^* \right\rangle$$

Combining with the two previous inequalities we get

$$\|x_t - x^*\|^2 \leq \|x_{t-1} - x^*\|^2 - 2\eta[f(x_t) - f(x^*)] + 2\eta^2\|\Delta_t\|^2 - 2\eta\left\langle \Delta_t, \bar{x}_t - x^* \right\rangle.$$

If we take the expectation with respect to i_t we have

$$\mathbb{E}\|x_t - x^*\|^2 \leq \|x_{t-1} - x^*\|^2 - 2\eta\mathbb{E}[f(x_t) - f(x^*)] + 2\eta^2\mathbb{E}\|\Delta_t\|^2 - 2\eta\left\langle \mathbb{E}\Delta_t, \bar{x}_t - x^* \right\rangle.$$

Now by using the Lemma 5 and $\|\bar{x}_t - x^*\| < Z$ we have

$$\begin{aligned}
&\mathbb{E}\|x_t - x^*\|^2 \\
&\leq \|x_{t-1} - x^*\|^2 - 2\eta\mathbb{E}[f(x_t) - f(x^*)] + 8\eta^2 L[f(x_{t-1}) - f(x^*) + f(\tilde{x}) - f(x^*)] + 2\eta^2\|e\|^2 + 2\eta\|e\|Z.
\end{aligned}$$

The rest of the proof follows the argument of [12], and is simlar to the previous proofs in this appendix. We take the expectation and sum up values, using convexity to give

$$2\eta(1 - 4L\eta)m[\mathbb{E}f(x^s) - f(x^*)] \leq (\frac{2}{\mu} + 8L\eta^2(m+1))[f(\tilde{x}_{s-1} - f(x^*)] + 2\eta^2\|e\|^2 + 2\eta\|e\|Z.$$

By dividing both sides to $2\eta(1 - 4L\eta)m$, we get the result. □

C.2 Proof of Proposition 3

We now turn to the case where h is differentiable, and we use an iteration that incorporates the gradient $h'(x_t)$. Recall that for this result we assume that g' is L_g-Lipschitz continuous, h' is L_h-Lipschitz continuous, and we defined $L_m = \max\{L_g, L_h\}$. If we let $\nu_t = h'(x_t) + g'_{i_t}(x_t) - g'_{i_t}(x^s) + \mu^s$, then note that we have $\mathbb{E}[\nu_t] = f'(x_t)$. Now as before we want to bound the expected second moment of ν_t,

$$\begin{aligned}
\mathbb{E}\left[\|\nu_t\|^2\right] &= \mathbb{E}\left[\|h'(x_t) + g'_{i_t}(x_t) - g'_{i_t}(x^s) + \mu^s\|^2\right] \\
&= \mathbb{E}[\|h'(x_t) + g'_{i_t}(x_t) - g'_{i_t}(x^s) + \mu^s + h'(x^*) + g'_{i_t}(x^*) - h'(x^*) - g'_{i_t}(x^*) + h'(x^s) - h'(x^s)\|^2] \\
&\leq 2\|h'(x_t) - h'(x^*)\|^2 + 2\mathbb{E}\left[\|g'_{i_t}(x_t) - g'_{i_t}(x^*)\|^2\right] \\
&\quad + 2\mathbb{E}\left[\| - g'_{i_t}(x^s) + \mu^s + h'(x^*) + g'_{i_t}(x^*) + h'(x^s) - h'(x^s)\|^2\right] \\
&= 2\|h'(x_t) - h'(x^*)\|^2 + 2\mathbb{E}\left[\|g'_{i_t}(x_t) - g'_{i_t}(x^*)\|^2\right] \\
&\quad + 2\mathbb{E}[\|g'_{i_t}(x^s) - \mu^s - h'(x^*) - g'_{i_t}(x^*) - h'(x^s) + h'(x^s) + h'(x^*) + g'(x^*)\|^2] \\
&= 2\|h'(x_t) - h'(x^*)\|^2 + 2\mathbb{E}\left[\|g'_{i_t}(x_t) - g'_{i_t}(x^*)\|^2\right] + 2\mathbb{E}\left[\|g'_{i_t}(x^s) - \mu^s - g'_{i_t}(x^*) + g'(x^*)\|^2\right] \\
&\leq 2\|h'(x_t) - h'(x^*)\|^2 + 2\mathbb{E}\left[\|g'_{i_t}(x_t) - g'_{i_t}(x^*)\|^2\right] + 2\mathbb{E}\left[\|g'_{i_t}(x^s) - g'_{i_t}(x^*)\|^2\right]
\end{aligned}$$

Now using that $\|h'(x^s) - h'(x^*)\|^2 \geq 0$ and $\|f(x) - f(y)\|^2 \leq 2L[f(x) - f(y) - \langle f'(y), x - y\rangle]$,

$$\begin{aligned}
\mathbb{E}\left[\|\nu_t\|^2\right] &\leq 2\|h'(x_t) - h'(x^*)\|^2 + 2\mathbb{E}\left[\|g'_{i_t}(x_t) - g'_{i_t}(x^*)\|^2\right] \\
&\quad + 2\mathbb{E}\left[\|g'_{i_t}(x^s) - g'_{i_t}(x^*)\|^2\right] + 2\|h'(x^s) - h'(x^*)\|^2 \\
&\leq 4L_h[h(x_t) - h(x^*) - \langle h'(x^*), x_t - x^*\rangle] + 4L_g[g(x_t) - g(x^*) - \langle g'(x^*), x_t - x^*\rangle] \\
&\quad + 4L_h[h(x^s) - h(x^*) - \langle h'(x^*), x^s - x^*\rangle] + 4L_g[g(x^s) - g(x^*) - \langle g'(x^*), x^s - x^*\rangle] \\
&\leq 4L_m[f(x_t) - f(x^*)] + 4L_m[f(x^s) - f(x^*)]
\end{aligned}$$

From this point, we follow the standard SVRG argument to obtain

$$\mathbb{E}\left[f(x_{s+1} - f(x^*)\right] \leq \left(\frac{1}{m\mu(1 - 2\eta L_m)} + \frac{2L_m\eta}{1 - 2\eta L_m}\right)[f(x_s - f(x^*)].$$

D Mini-Batch

We first give an analysis of SVRG where mini-batches are selected by sampling propotional to the Lipschitz constants of the gradients. We then consider the mixed deterministic/random sampling scheme described in the main paper.

D.1 SVRG with Mini-batch

Here we consider using a 'mini-batch' of examples in the *inner* SVRG loop. We use M to denote the batch size, and we assume that the elements of the mini-batch are sampled with a probability of $p_i = L_i/n\bar{L}$. This gives a search direction and inner iteration of:

$$\begin{aligned}
\nu_t &= \mu^s + \frac{1}{M}\left[\sum_{i \in M} \frac{1}{np_i}\left(f'_i(x_t) - f'_i(x^s)\right)\right], \\
x_{t+1} &= x_t - \eta\nu_t.
\end{aligned}$$

Observe that $\mathbb{E}[\nu_t] = f'(x_t)$, and since each f_i is L_i-smooth we still have that

$$\|f'_i(x) - f'_i(y)\|^2 \leq L_i\left(f_i(x) - f_i(y) - \langle f'_i(y), x - y\rangle\right), .$$

It follows from the definition of p_i that

$$\mathbb{E}\left[\left\|\frac{1}{np_i}f_i'(x) - f_i'(y)\right\|^2\right] = \frac{1}{n}\sum_i \frac{1}{np_i}\|f_i'(x) - f_i'(y)\|^2 \\ \leq 2\bar{L}\left(f(x) - f(y) - \langle f'(y), x - y\rangle\right),$$

which we use to bound $\mathbb{E}\left[\|\nu_t\|^2\right]$ as before,

$$\begin{aligned}
\mathbb{E}\left[\|\nu_t\|^2\right] &= \mathbb{E}\left[\left\|\frac{1}{M}\sum_i\left(\frac{1}{np_i}(f_i'(x_t) - f_i'(x^s) + \mu^s\right)\right\|^2\right] \\
&= \mathbb{E}\left[\left\|\frac{1}{M}\sum_i\left(\frac{1}{np_i}(f_i'(x_t) - f_i'(x^*) + f_i'(x^*) - f_i'(x^s) + \mu^s\right)\right\|^2\right] \\
&\leq \frac{2}{M^2}\sum_i \mathbb{E}\left[\left\|\left(\frac{1}{np_i}(f_i'(x_t) - f_i'(x^*))\right)\right\|^2\right] \\
&\quad + \frac{2}{M^2}\sum_i \mathbb{E}\left[\left\|\left(\frac{1}{np_i}(f_i'(x^s) - f_i'(x^*)) - \mu^s\right)\right\|^2\right] \\
&\leq \frac{2}{M^2}\sum_i \mathbb{E}\left[\left\|\left(\frac{1}{np_i}(f_i'(x_t) - f_i'(x^*))\right)\right\|^2\right] \\
&\quad + \frac{2}{M^2}\sum_i \mathbb{E}\left[\left\|\left(\frac{1}{np_i}(f_i'(x^s) - f_i'(x^*))\right)\right\|^2\right] \\
&\leq \frac{4\bar{L}}{M}\left[f(x_t) - f(x^*)\right] + \frac{4\bar{L}}{M}\left[f(x^s) - f(x^*)\right]
\end{aligned}$$

It subsequently follows that

$$\mathbb{E}\left[f(x^{s+1}) - f(x^*)\right] \leq \left(\frac{M}{m\mu(M - 2\eta\bar{L})\eta} + \frac{2\bar{L}\eta}{M - 2\eta\bar{L}}\right)\mathbb{E}\left[f(x^s) - f(x^*)\right]$$

D.2 Proof of Proposition 4

We now consider the case where we have $g(x) = (1/n)\sum_{i\notin[\mathcal{B}_f]} f_i(x)$ and $h(x) = (1/n)\sum_{i\in[\mathcal{B}_f]} f_i(x)$ for some batch $\mathcal{B}_f$. We assume that we sample M_r elements of g with probability of $p_i = \frac{L_i}{(n-M_f)\bar{L}_r}$ and that we use:

$$\begin{aligned}
\nu_t &= g'(x^s) + h'(x_t) + \frac{1}{M_r}\left[\sum_{i\in M_r}\frac{1}{np_i}\left(f_i'(x_t) - f_i'(x^s)\right)\right], \\
&= \mu^s + h'(x_t) + \frac{1}{M_r}\left[\sum_{i\in M_r}\frac{1}{np_i}\left(f_i'(x_t) - f_i'(x^s)\right)\right] - h'(x^s), \\
x_{t+1} &= x_t - \eta\nu_t,
\end{aligned}$$

where as usual $\mu^s = \frac{1}{n}\sum_{i=1}^n f_i'(x^s) = g'(x^s) + h'(x^s)$. Note that $\mathbb{E}\left[\nu_t\right] = f'(x_t)$. We first bound $\mathbb{E}\left[\|\nu_t\|^2\right]$,

$$\begin{aligned}
\mathbb{E}\left[\|\nu_t\|^2\right] &= \mathbb{E}\left[\left\|\mu^s + h'(x_t) + 1/M_r\left[\sum_{i\in M_r}\frac{1}{np_i}\left(f_i'(x_t) - f_i'(x^s)\right)\right] - h'(x^s)\right\|^2\right] \\
&= \mathbb{E}[\|\mu^s + h'(x_t) - h'(x^*) + 1/M_r\left[\sum_{i\in M_r}\frac{1}{np_i}\left(f_i'(x_t) - f_i'(x^*)\right)\right] \\
&- 1/M_r\left[\sum_{i\in M_r}\frac{1}{np_i}\left(f_i'(x^s) - f_i'(x^*)\right)\right] - h'(x^s) + h'(x^*)\|^2] \\
&\leq 2/n^2\underbrace{\sum_{i\in\mathcal{B}_f}\|f_i'(x_t) - f_i'(x^*)\|^2}_{\text{Fixed part}} + \underbrace{2/M_r^2\sum_{i\in\mathcal{B}_r}\mathbb{E}\left[\left\|\frac{1}{np_i}(f_i'(x_t) - f_i'(x^*))\right\|^2\right]}_{\text{Random part}} \\
&+ 2\mathbb{E}\left[\left\|1/M_r\left[\sum_{i\in M_r}\frac{1}{np_i}\left(f_i'(x^s) - f_i'(x^*)\right)\right] + h'(x^s) - h'(x^*) - \mu^s\right\|^2\right],
\end{aligned}$$

where the inequality uses $\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$. Now we bound each of the above terms separately,

$$\begin{aligned}
2/n^2\sum_{i\in\mathcal{B}_f}\|f_i'(x_t) - f_i'(x^*)\|^2 &\leq 2/n^2\sum_{i\in\mathcal{B}_f}2L_i(f_i(x_t) - f_i(x^*) - \langle f_i'(x^*), x_t - x^*\rangle) \\
&\leq 4L_1/n\left(h(x_t) - h(x^*) - \langle h'(x^*), x_t - x^*\rangle\right),
\end{aligned}$$

$$\begin{aligned}
&2/M_r^2\sum_{i\in\mathcal{B}_r}\mathbb{E}\left[\|\frac{1}{np_i}(f_i'(x_t) - f_i'(x^*))\|^2\right] \\
&\quad\leq 2/M_r^2\sum_{i\in\mathcal{B}_r}1/n^2\sum_{j\notin\mathcal{B}_f}1/p_i\|f_i'(x_t) - f_i'(x^*)\|^2 \\
&\quad\leq 2/M_r^2\sum_{i\in\mathcal{B}_r}1/n^2\sum_{j\notin\mathcal{B}_f}(n - M_f)\bar{L}_r\left(f_i(x_t) - f_i(x^*) - \langle f_i'(x^*), x_t - x^*\rangle\right) \\
&\quad= \frac{4(n-M_f)\bar{L}_r}{nM_r}(g(x_t) - g(x^*) - \langle g'(x^*), x_t - x^*\rangle).
\end{aligned}$$

Finally for the last term we have,

$$\begin{aligned}
&2\mathbb{E}\left[\left\|\frac{1}{M_r}\left[\sum_{i\in M_r}\frac{1}{np_i}\left(f_i'(x^s) - f_i'(x^*)\right)\right] + \underbrace{h'(x^s) - h'(x^*) - \mu^s}_{=g'(x^*) - g'(x^s)}\right\|^2\right] \\
&\leq 2\mathbb{E}\left[\left\|\frac{1}{M_r}\sum_{i\in M_r}\frac{1}{np_i}\left(f_i'(x^s) - f_i'(x^*)\right)\right\|^2\right] \\
&\leq \frac{4(n-M_f)\bar{L}_r}{nM_r}\left(g(x^s) - g(x^*) - \langle g'(x^*), x^s - x^*\rangle\right)
\end{aligned}$$

where the first inequality uses variance inequality ($\mathbb{E}\|X - \mathbb{E}X\|^2 \leq \mathbb{E}\|X\|^2$) and the second one comes from Lemma 1. Since h is convex we can add $\frac{4(n-M_f)\bar{L}_r}{nM_r}(h(x^s) - h(x^*) - \langle h'(x^*), x^s - x^*\rangle)$ to the right side of the above term,

giving

$$2\mathbb{E}\left[\left\|\frac{1}{M_r}\left[\sum_{i\in M}\frac{1}{np_i}\left(f_i'(x^s)-f_i'(x^*)\right)\right]+h'(x^s)-h'(x^*)-\mu^s\right\|^2\right]$$
$$\leq \frac{4(n-M_f)\bar{L}_r}{nM_r}\left(f(x^s)-f(x^*)\right).$$

Now following the proof technique we used several times, we can show that:

$$\mathbb{E}\left[f(x^{s+1})-f(x^*)\right]\leq\left(\frac{1}{m\mu(1-2\eta\kappa)\eta}+\frac{2\zeta\eta}{1-2\eta\kappa}\right)\mathbb{E}\left[f(x^s)-f(x^*)\right]$$

where $\zeta=\frac{(n-M_f)\bar{L}_r}{(M-M_f)n}$ and $\kappa=\max\{\frac{L_1}{n},\zeta\}$.

E Learning efficiency

In this section we closely follow Bottou and Bousquet [1, 2] to discuss the performance of SVRG, and other linearly-convergent stochastic methods, as learning algorithms. In the typical supervised learning setting, we are giving n independently drawn input-output pairs (x_i, y_i) from some distribution $P(x, y)$ and we seek to minimize the empirical risk,

$$E_n(f)=\frac{1}{n}\sum_{i=1}^{n}\ell(f(x_i),y_i)=\mathbb{E}_n[\ell(f(x),y)],$$

where ℓ is our loss function. However, in machine learning this is typically just a surrogate for the objective we are ultimately interested in. In particular, we typically want to minimize the *expected* risk,

$$E(f)=\int\ell(f(x),y)dP(x,y)=\mathbb{E}[\ell(f(x),y)],$$

which tells us how well we do on test data from the same distribution. We use f^* to denote the minimizer of the expected risk,

$$f^*(x)=\arg\min_{\hat{y}}\mathbb{E}[\ell(\hat{y},y)\,|\,x],$$

which is the best that a learner can hope to achieve.

Consider a family $\mathcal{F}$ of possible functions that we use to predict y_i from x_i. We write the minimizer of the expected risk over this restricted set as $f^*_{\mathcal{F}}$,

$$f^*_{\mathcal{F}}=\arg\min_{f\in\mathcal{F}}E(f),$$

while we denote the empirical risk minimizer within this family as f_n,

$$f_n=\arg\min_{f\in\mathcal{F}}E_n(f).$$

But, since we are applying a numerical optimizer we only assume that we find a ρ-optimal minimizer of the empirical risk $\tilde{f}_n$,

$$E_n(\tilde{f}_n)<E_n(f_n)+\rho,$$

In this setting, Bottou & Bousquet consider writing the sub-optimality of the approximate empirical risk minimizer $\tilde{f}_n$ compared to the minimizer of the expected risk f^* as

$$\begin{aligned}\mathcal{E}&=\mathbb{E}[E(\tilde{f}_n)-E(f^*)]\\&=\underbrace{\mathbb{E}[E(f^*_{\mathcal{F}})-E(f^*)]}_{\mathcal{E}_{\text{app}}}+\underbrace{\mathbb{E}[E(f_n)-E(f^*_{\mathcal{F}})]}_{\mathcal{E}_{\text{est}}}+\underbrace{\mathbb{E}[E(\tilde{f}_n)-E(f_n)]}_{\mathcal{E}_{\text{opt}}},\end{aligned}\tag{1}$$

where the expectation is taken with respect to the output of the algorithm and with respect to the training examples that we sample. This decomposition shows how three intuitive terms affect the sub-optimality:

1. $\mathcal{E}_{\text{app}}$ is the *approximation error*: it measures the effect of restricting attention to the function class $\mathcal{F}$.

2. $\mathcal{E}_{\text{est}}$ is the *estimation error*: it measures the effect of only using a finite number of samples.

3. $\mathcal{E}_{\text{opt}}$ is the *optimization error*: it measures the effect of inexactly solving the optimization problem.

While choosing the family of possible approximating functions $\mathcal{F}$ is an interesting and important issue, for the remainder of this section we will assume that we are given a fixed family. In particular, Bottou & Bousquet's assumption is that $\mathcal{F}$ is linearly-parameterized by a vector $w \in \mathbb{R}^d$, and that all quantities are bounded (x_i, y_i, and w). This means that the approximation error $\mathcal{E}_{\text{app}}$ is fixed so we can only focus on the trade-off between the estimation error $\mathcal{E}_{\text{est}}$ and the optimization error $\mathcal{E}_{\text{opt}}$.

All other sections of this work focus on the case of *finite* datasets where we can afford to do several passes through the data (*small-scale learning problems* in the language of Bottou & Bousquet). In this setting, $\mathcal{E}_{\text{est}}$ is fixed so all we can do to minimize $\mathcal{E}$ is drive the optimization error ρ as small as possible. In this section we consider the case where we do not have enough time to process all available examples, or we have an infinite number of possible examples (*large-scale learning problems* in the language of Bottou & Bousquet). In this setting, the time restriction means we need to make a trade-off between the optimization error and the estimation error: should we increase n in order to decrease the estimation error $\mathcal{E}_{\text{est}}$ or should we revisit examples to try to more quickly decrease the optimization error $\mathcal{E}_{\text{opt}}$ while keeping the estimation error fixed?

Bottou & Bousquet discuss how under various assumptions we have the variance condition

$$\forall f \in \mathcal{F} \quad \mathbb{E}\left[\left(\ell(f(X),Y) - \ell(f^*_{\mathcal{F}}(X),Y)\right)^2\right] \leq c\left(E(f) - E(f^*_{\mathcal{F}})\right)^{2-\frac{1}{\alpha}},$$

and how this implies the bound

$$\mathcal{E} = O\left(\mathcal{E}_{\text{app}} + \left(\frac{d}{n}\log\frac{n}{d}\right)^{\alpha} + \rho\right).$$

To make the second and third terms comparable, we can take $\rho = \left(\frac{d}{n}\log\frac{n}{d}\right)^{\alpha}$. Then to achieve an accuracy of $O(\mathcal{E}_{\text{app}} + \epsilon)$ it is sufficient to take $n = O\left(\frac{d}{\epsilon^{1/\alpha}}\log(1/\epsilon)\right)$ samples:

$$\begin{aligned}
\mathcal{E} &= O\left(\mathcal{E}_{\text{app}} + \left(\frac{d}{n}\log\frac{n}{d}\right)^{\alpha} + \rho\right) \\
&= O\left(\mathcal{E}_{\text{app}} + \left(\frac{d}{n}\log\frac{n}{d}\right)^{\alpha} + \left(\frac{d}{n}\log\frac{n}{d}\right)^{\alpha}\right) \\
&= O\left(\mathcal{E}_{\text{app}} + \left(\frac{d}{n}\log\frac{n}{d}\right)^{\alpha}\right) \\
&= O\left(\mathcal{E}_{\text{app}} + \left(\frac{\epsilon^{\frac{1}{\alpha}}}{\log(\frac{1}{\epsilon})}\log\left(\frac{\log(\frac{1}{\epsilon})}{\epsilon^{\frac{1}{\alpha}}}\right)\right)^{\alpha}\right) \\
&= O\left(\mathcal{E}_{\text{app}} + \epsilon\left(\frac{\log(\log(1/\epsilon)) - \frac{1}{\alpha}\log(\epsilon)}{\log(1/\epsilon)}\right)^{\alpha}\right) \\
&= O(\mathcal{E}_{\text{app}} + \epsilon).
\end{aligned}$$

The results presented in the main paper follow from noting that (i) the iteration cost of SVRG is $O(d)$ and (ii) that the number of iterations for SVRG to reach an accuracy of ρ is $O((n+\kappa)\log(1/\rho))$.

F Additional Experimental Results

We list properties of the dataset considered in the experiments in Table 1. In Figures 1-4, we plot the performance on the various datasets in terms of both the training objective and test error, showing the maximum/mean/minimum performance across 10 random trials. In these plots, we see a clear advantage for the *Grow* strategy on the largest

Data set	Data Points	Variables	Reference
quantum	50 000	78	[4]
protein	145 751	74	[4]
sido	12 678	4 932	[7]
rcv1	20 242	47 236	[10]
covertype	581 012	54	[6]
news	19 996	1 355 191	[9]
spam	92 189	823 470	[5, 3]
rcv1Full	697 641	47 236	[10]
alpha	500 000	500	Synthetic

Table 1: Binary data sets used in the experiments.

datasets (bottom row), but less of an advantage or no advantage on the smaller datasets. The advantage of using support vectors seemed less dependent on the data size, as it helped in some small datasets as well as some large datasets, while in some small/large datasets it did not make a big difference.

In Figures 5-6, we give the result of experiments comparing different mini-batch selection strategies. In particular, we consider mini-batch SVRG with a batch size of 16 and compare the following methods: uniform sampling of the mini-batch (*Uniform*), sampling proportional to the Lipschitz constants (*Lipschitz*), and a third strategy based on Proposition 4 in the main paper (*Lipschitz+*). On each iteration, the *Lipschitz+* strategy constructs the mini-batch using the 100 examples with the largest Lipschitz constants (the 'fixed' set) in addition to 16 examples sampled according to their Lipschitz constants from among the remaining examples. We assume that the fixed set is computed 'for free' by calculating these gradients on a GPU or FPGA. In these experiments, there was often no difference between the various methods because the rows of the data were normalized. For the two Lipschitz sampling strategies, we used a step size of $1/\bar{L}$. In some cases, the new sampling scheme may have given a small improvement, but in general the theoretical advantage of this method was not reflected in our experiments.

In Figure 7-8, we repeat the mini-batch experiment but include two additional method: sampling example i proportional to $f_i(x^s)$ (*Function*) and sampling i proportional to $\|f_i'(x^s)\|$ (*Gradient*). For these strategies we used a step size of $1/\bar{L}$, and on eight of the nine datasets we were surprised that these strategies had similar performance to the Lipschitz sampling strategy (even though they do not have access to the L_i). However, both of these strategies had strange behaviour on one of the datasets. On the *covertype* dataset, the *Function* method seemed to diverge in terms of training objective and test error while the *Gradient* seemed to converge to a sub-optimal solution in terms of training objective but achieved close to the optimal test error.

References

[1] L. Bottou and O. Bousquet. The tradeoffs of large scale learning. *Advances in neural information processing systems (NIPS)*, 2007.

[2] Léon Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, pages 177–186. Springer, 2010.

[3] P. Carbonetto. *New probabilistic inference algorithms that harness the strengths of variational and Monte Carlo methods*. PhD thesis, Univ. of British Columbia, May 2009.

[4] R. Caruana, T. Joachims, and L. Backstrom. KDD-cup 2004: results and analysis. *ACM SIGKDD Newsletter*, 6(2):95–108, 2004.

[5] G. V. Cormack and T. R. Lynam. Spam corpus creation for TREC. In *Proc. 2nd Conference on Email and Anti-Spam*, 2005. `http://plg.uwaterloo.ca/~gvcormac/treccorpus/`.

[6] A. Frank and A. Asuncion. UCI machine learning repository, 2010.

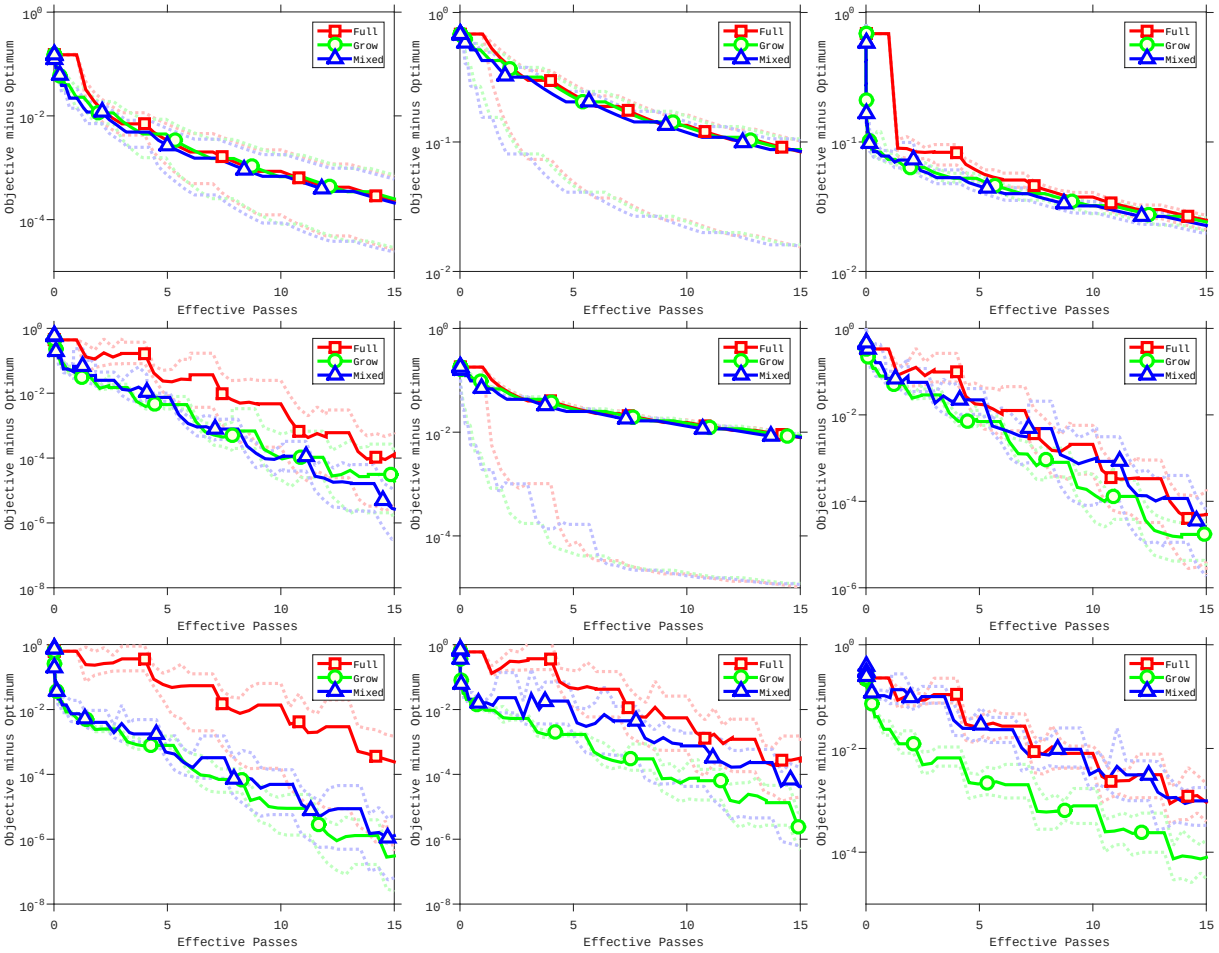

Figure 1: Comparison of training objective of logistic regression for different datasets. The top row gives results on the *quantum* (left), *protein* (center) and *sido* (right) datasets. The middle row gives results on the *rcv11* (left), *covertype* (center) and *news* (right) datasets. The bottom row gives results on the *spam* (left), *rcv1Full* (center), and *alpha* (right) datasets.

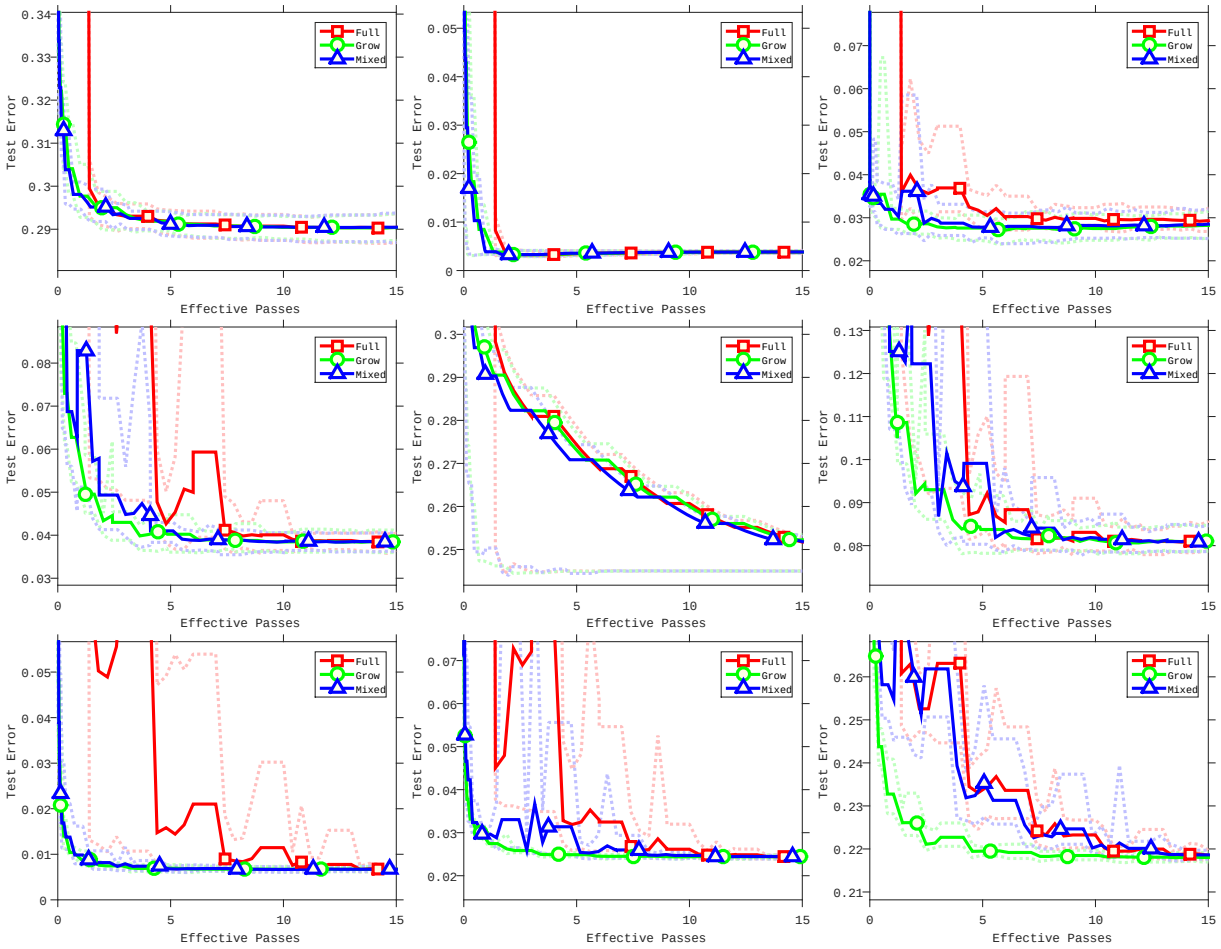

Figure 2: Comparison of test error of logistic regression for different datasets. The top row gives results on the *quantum* (left), *protein* (center) and *sido* (right) datasets. The middle row gives results on the *rcv11* (left), *covertype* (center) and *news* (right) datasets. The bottom row gives results on the *spam* (left), *rcv1Full* (center), and *alpha* (right) datasets.

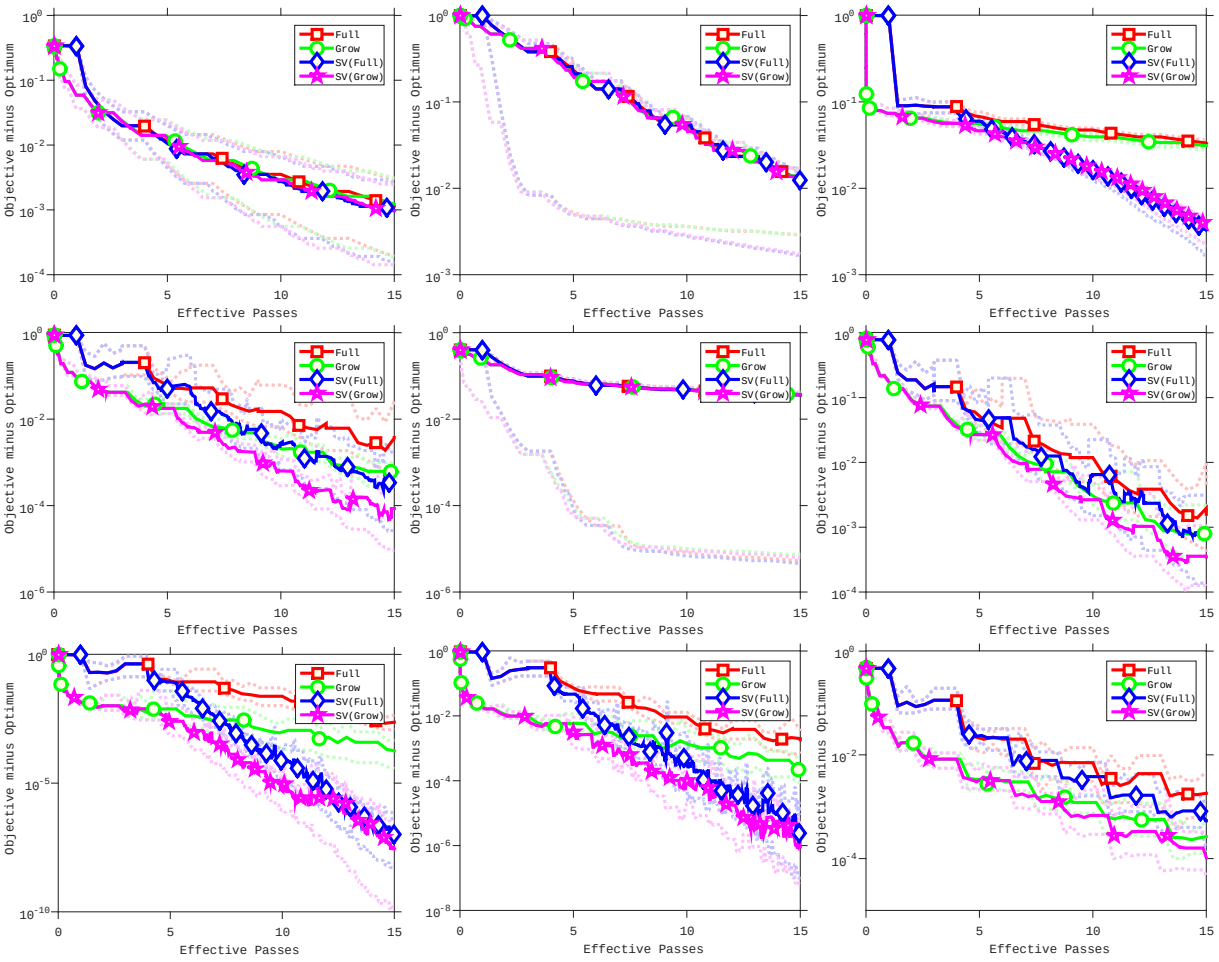

Figure 3: Comparison of training objective of SVM for different datasets. The top row gives results on the *quantum* (left), *protein* (center) and *sido* (right) datasets. The middle row gives results on the *rcv11* (left), *covertype* (center) and *news* (right) datasets. The bottom row gives results on the *spam* (left), *rcv1Full* (center), and *alpha* (right) datasets.

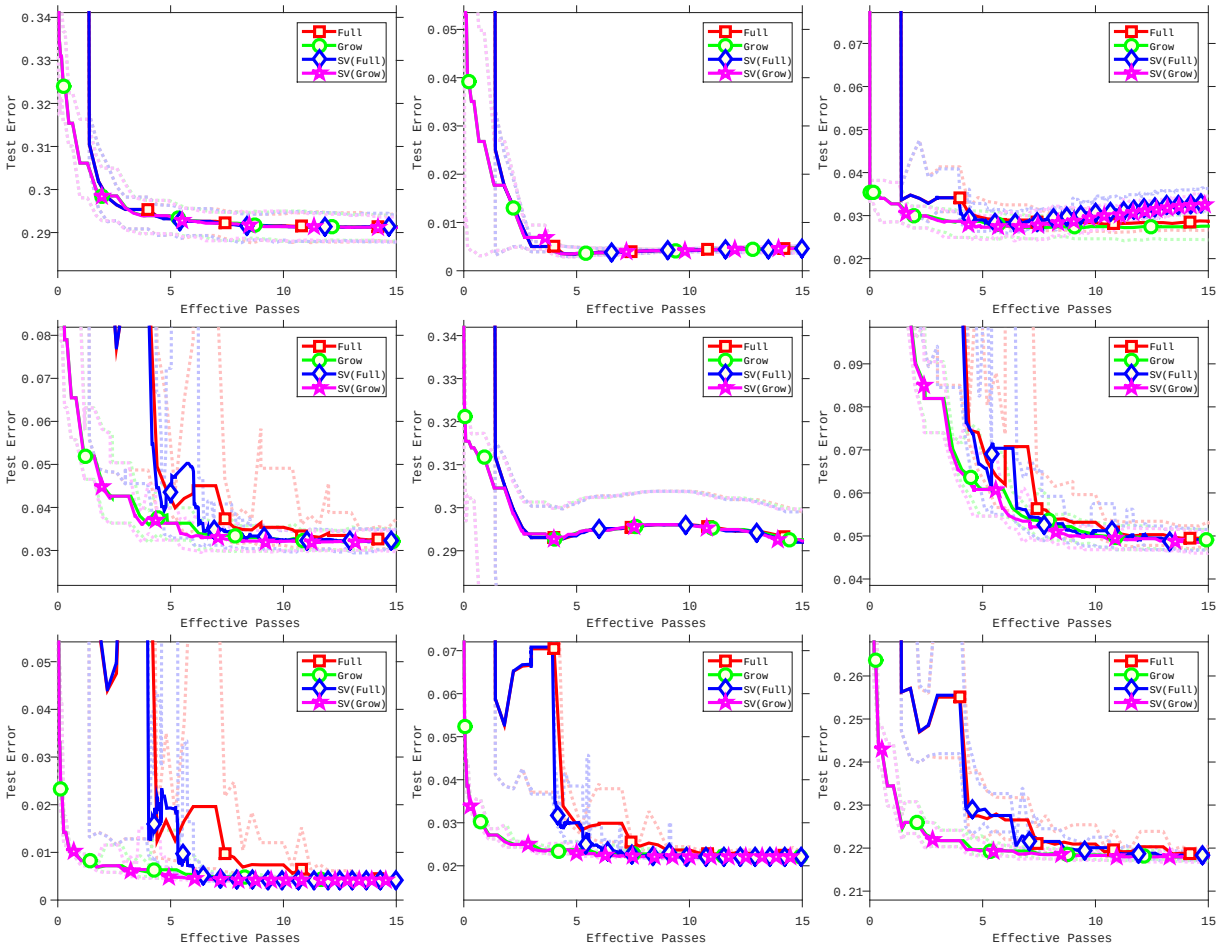

Figure 4: Comparison of test error of SVM for different datasets. The top row gives results on the *quantum* (left), *protein* (center) and *sido* (right) datasets. The middle row gives results on the *rcv11* (left), *covertype* (center) and *news* (right) datasets. The bottom row gives results on the *spam* (left), *rcv1Full* (center), and *alpha* (right) datasets.

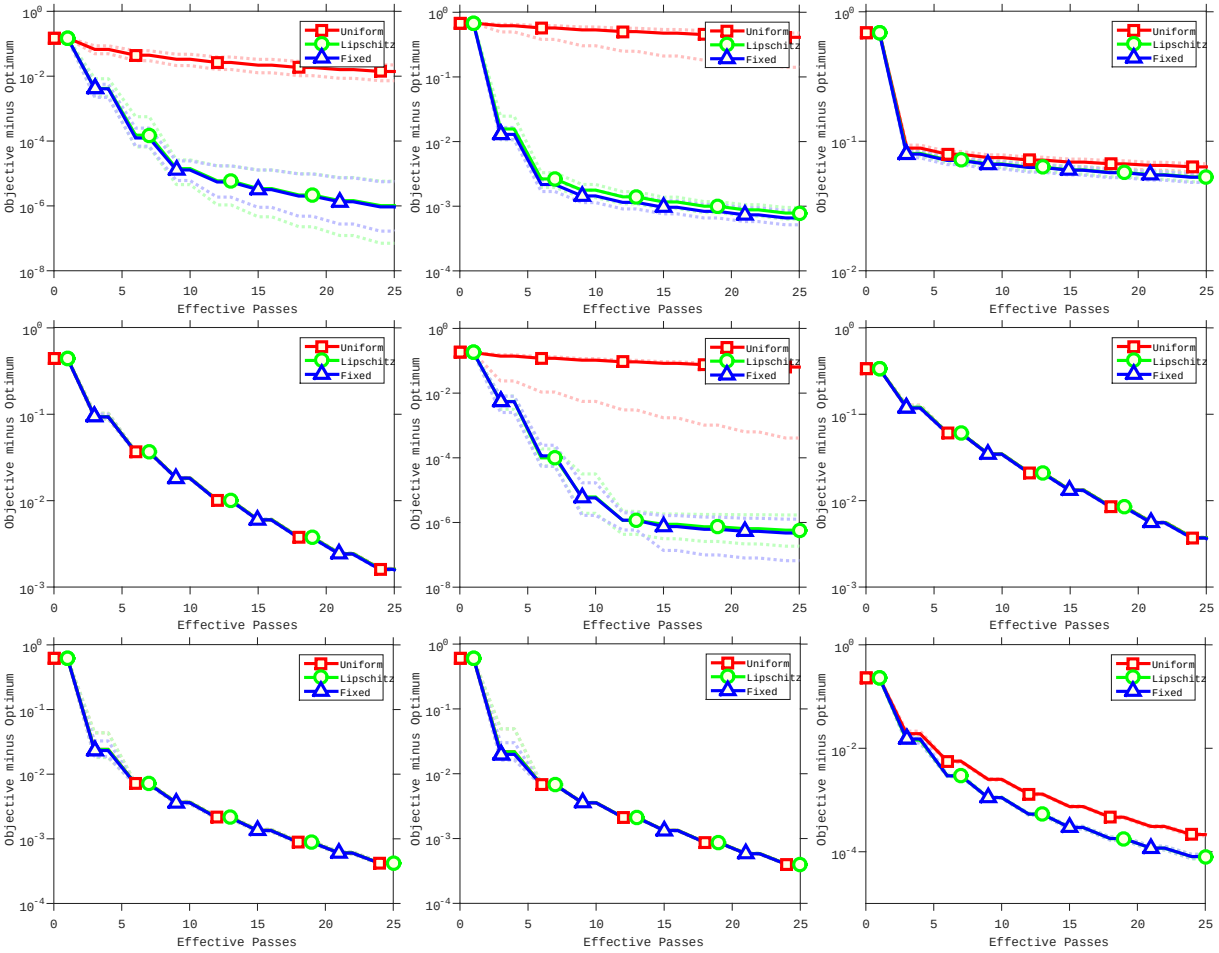

Figure 5: Comparison of training objective of logistic regression with different mini-batch strategies. The top row gives results on the *quantum* (left), *protein* (center) and *sido* (right) datasets. The middle row gives results on the *rcv11* (left), *covertype* (center) and *news* (right) datasets. The bottom row gives results on the *spam* (left), *rcv1Full* (center), and *alpha* (right) datasets.

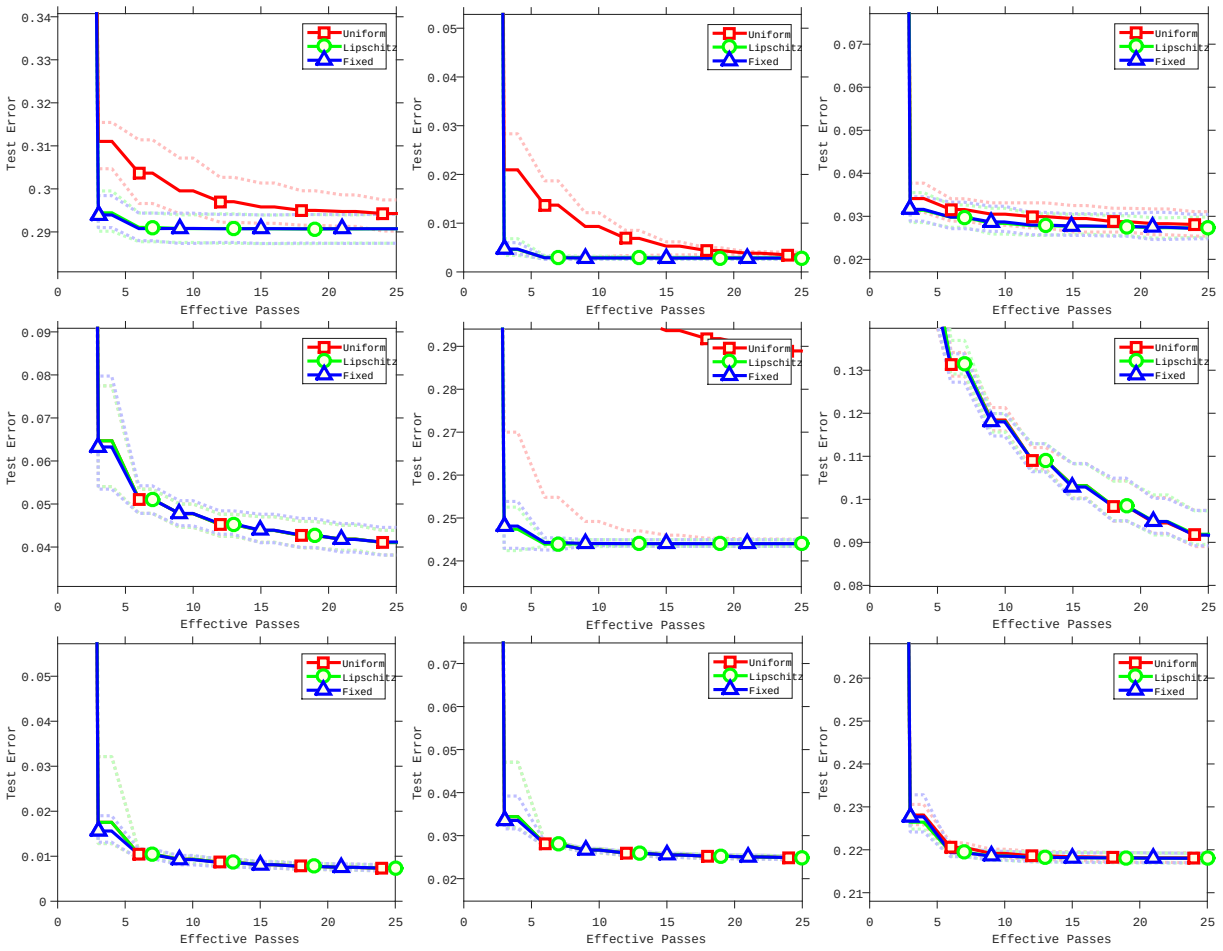

Figure 6: Comparison of test error of logistic regression with different mini-batch strategies. The top row gives results on the *quantum* (left), *protein* (center) and *sido* (right) datasets. The middle row gives results on the *rcv11* (left), *covertype* (center) and *news* (right) datasets. The bottom row gives results on the *spam* (left), *rcv1Full* (center), and *alpha* (right) datasets.

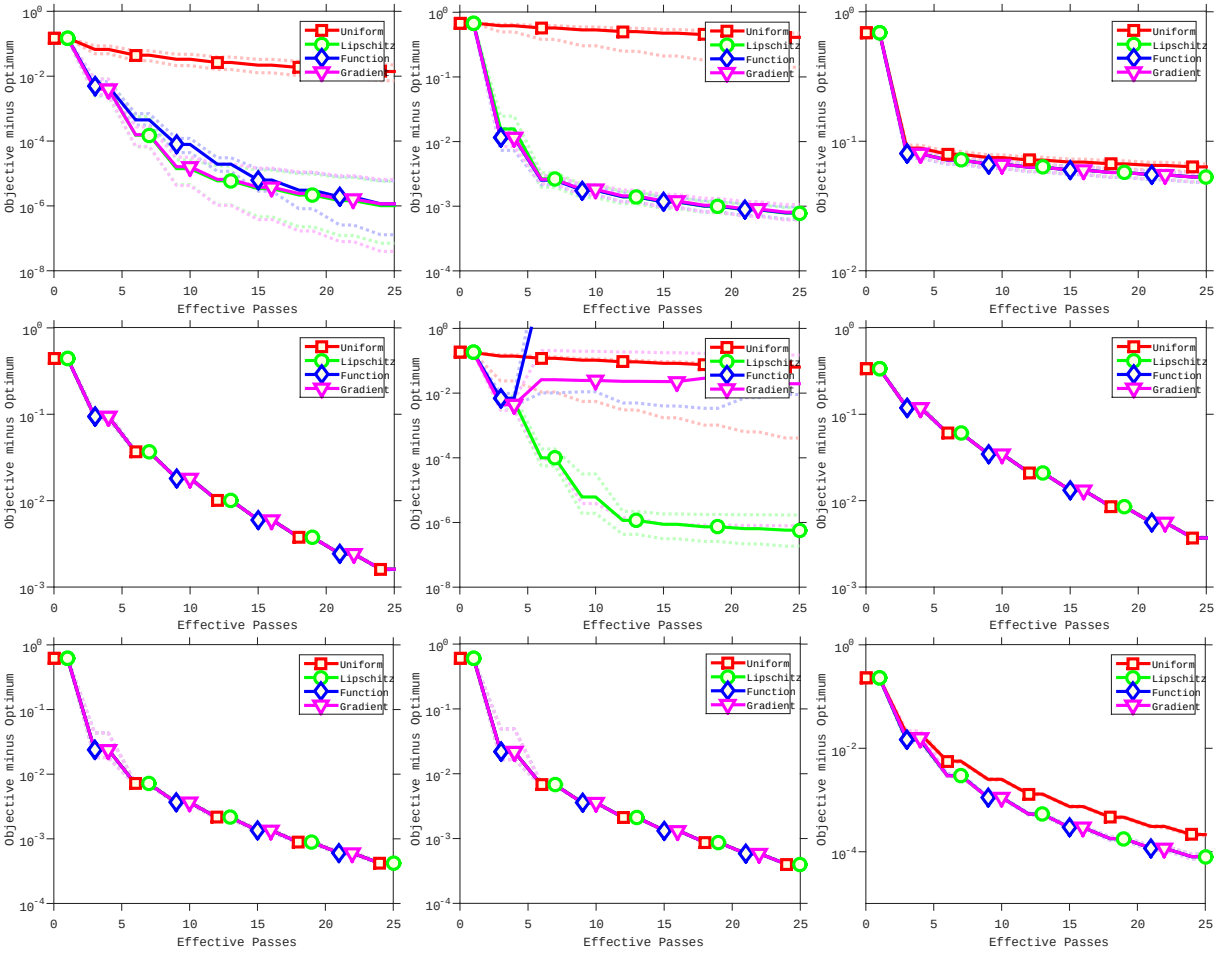

Figure 7: Comparison of training objective of logistic regression with different mini-batch strategies. The top row gives results on the *quantum* (left), *protein* (center) and *sido* (right) datasets. The middle row gives results on the *rcv11* (left), *covertype* (center) and *news* (right) datasets. The bottom row gives results on the *spam* (left), *rcv1Full* (center), and *alpha* (right) datasets.

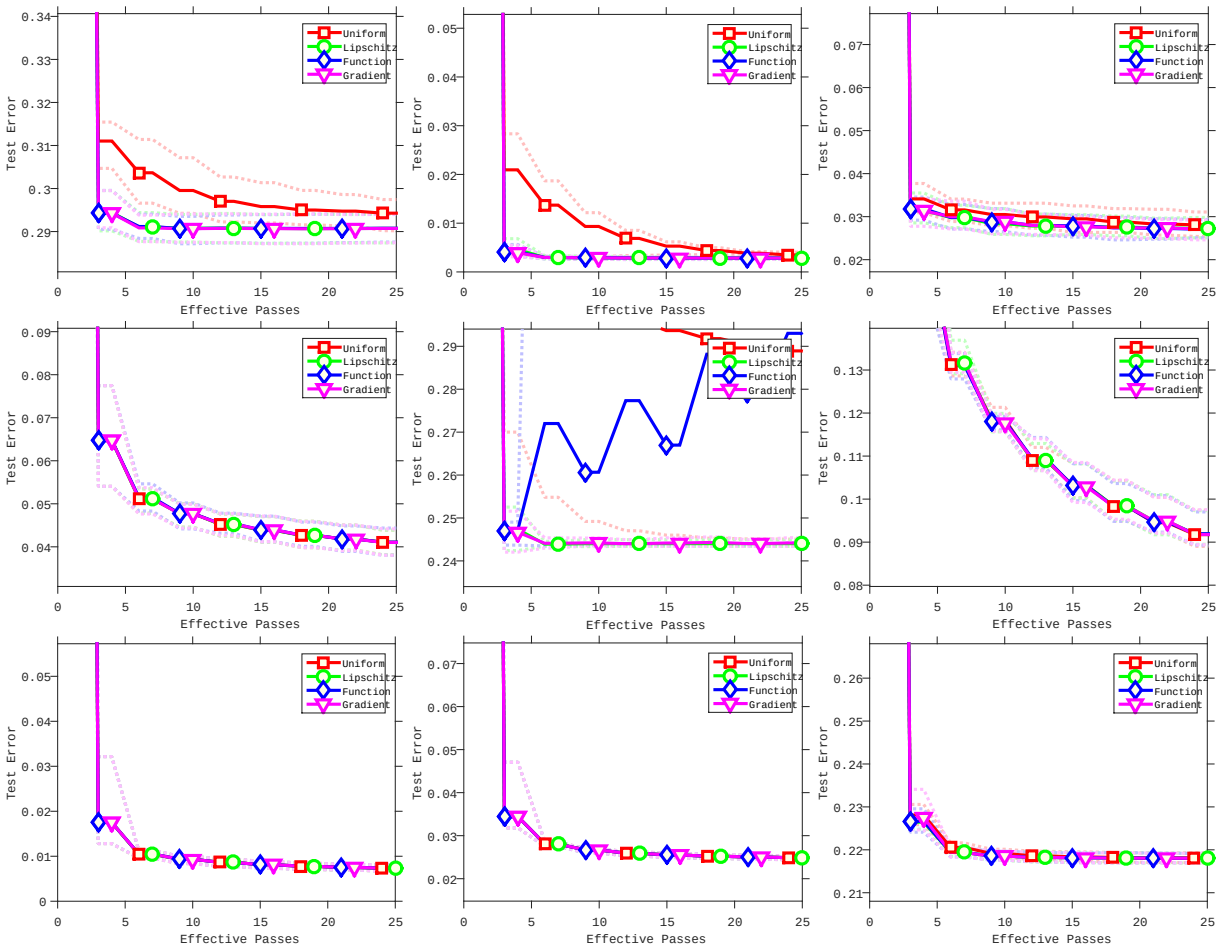

Figure 8: Comparison of test error of logistic regression with different mini-batch strategies. The top row gives results on the *quantum* (left), *protein* (center) and *sido* (right) datasets. The middle row gives results on the *rcv11* (left), *covertype* (center) and *news* (right) datasets. The bottom row gives results on the *spam* (left), *rcv1Full* (center), and *alpha* (right) datasets.

[7] I. Guyon. Sido: A phamacology dataset, 2008.

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

[9] S.S. Keerthi and D. DeCoste. A modified finite newton method for fast solution of large scale linear svms. *Journal of Machine Learning Research*, 6:341–361, 2005.

[10] D.D. Lewis, Y. Yang, T. Rose, and F. Li. RCV1: A new benchmark collection for text categorization research. *Journal of Machine Learning Research*, 5:361–397, 2004.

[11] Y. Nesterov. *Introductory lectures on convex optimization: A basic course*. Springer, 2004.

[12] Lin Xiao and Tong Zhang. A proximal stochastic gradient method with progressive variance reduction. *SIAM Journal on Optimization*, 24(2):2057–2075, 2014.