Stochastic Proximal Gradient Method: Variance Reduction and Support Identification

Yutong Dai [1] Guanyi Wang[2] Frank E. Curtis[1] Daniel P. Robinson[1]

Lehigh University[1] National University of Singapore[2]

INFORMS Annual Meeting 2023

Oct 16, 2023

Outline

Outline

Problem of Interest

Sparse optimization problem

$$\min_{x\in\mathbb{R}^n} \quad f(x) + r(x) := \mathbb{E}_{\xi\sim\mathcal{P}}[\ell(x;\xi)] + r(x)$$

- $\ell(x;\xi)$: loss function; convex and smooth almost surely.
  - regression problem: $f(x) = \frac{1}{N}\sum_{i=1}^{N} \log(1+e^{-y_i x^T d_i})$
  - diffusion problem: $f(x) = \mathbb{E}_{t\sim[1,T],p\ \sim\mathcal{P},\epsilon_t\sim N(0,I)}\left[\left\|\epsilon_t - \ell\left(x; \sqrt{\bar{\alpha}_t}p + \sqrt{1-\bar{\alpha}_t}\epsilon_t, t\right)\right\|^2\right]$
- $r(x)$: group sparsity inducing regularizer; convex and nonsmooth:
  - group ℓ_1: $r(x) = \sum_{i\in n_{\mathcal{G}}} \lambda_i \|[x]_{g_i}\|_2 \quad \left(\lambda_i > 0 \text{ for all } i \in n_{\mathcal{G}} \text{ and } \bigcup_{i\in n_{\mathcal{G}}} g_i = [n]\right)$
  - Example: for $x \in \mathbb{R}^3$

non-overlapping $\quad g_1 = \{1,2\}$ and $g_2 = \{3\}$: $r(x) = \lambda_1 \left\|\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}\right\| + \lambda_2\|x_3\|.$

overlapping $\quad g_1 = \{1,2\}$ and $g_2 = \{2,3\}$: $r(x) = \lambda_1 \left\|\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}\right\| + \lambda_2 \left\|\begin{pmatrix} x_2 \\ x_3 \end{pmatrix}\right\|.$

(Stochastic) Proximal Gradient Methods

- **Access to true gradients**

$$x_{k+1} = \mathrm{Prox}_{\alpha_k r}(x_k - \alpha_k \nabla f(x_k)) := \arg\min_{x\in\mathbb{R}^n} \left\{ \frac{1}{2\alpha_k}\|x - (x_k - \alpha_k \nabla f(x_k))\|_2^2 + r(x) \right\}$$

- **No/Restricted access to true gradients**

$$x_{k+1} = \mathrm{Prox}_{\alpha_k r}(x_k - \alpha_k d_k) := \arg\min_{x\in\mathbb{R}^n} \left\{ \frac{1}{2\alpha_k}\|x - (x_k - \alpha_k d_k)\|_2^2 + r(x) \right\}$$

with d_k being some form of stochastic gradient estimator for $\nabla f(x_k)$, e.g.,

$$d_k = \nabla \ell(x_k; \xi) \text{ with } \xi \sim \mathcal{P}.$$

Variance Reduction

finite-sum structure

$f(x) = \frac{1}{N}\sum_{i=1}^{N} \ell(x;\xi_i)$. variance reduced d_k is constrcuted by using the control of variate idea.

- SAGA[1]: form a gradient table $G = [\nabla\ell(x_{t_1};\xi_1), \cdots, \nabla\ell(x_{t_N};\xi_N)] \in \mathbb{R}^{n\times N}$,

$$d_k = \nabla\ell(x_k;\xi_i) - G[:,i] + \frac{1}{N}G\mathbf{1} \text{ and } G[:,i] \leftarrow \nabla\ell(x_k;\xi_i)$$

- ProxSVRG[2]: periodic full gradient evaluation at anchor point $\tilde{x}_k$.

$$d_k = \nabla\ell(x_k;\xi_i) - \nabla\ell(\tilde{x}_k;\xi_i) + \nabla f(\tilde{x}_k) \text{ and } \tilde{x}_k \text{ is updated periodically}$$

Other methods ProxSARAH, ProxSpider, and more ...

[1] Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. "SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives". In: *Advances in neural information processing systems* 27 (2014).

[2] Lin Xiao and Tong Zhang. "A proximal stochastic gradient method with progressive variance reduction". In: *SIAM Journal on Optimization* 24.4 (2014), pp. 2057–2075.

Support Identification

The **support** of a point $x \in \mathbb{R}^n$ is defined as

$$\mathcal{S}(x) = \{i \in \{1, \ldots, n_{\mathcal{G}}\} \mid [x]_{g_i} \neq 0\}.$$

We say that **support identification** happens at point $x \in \mathbb{R}^n$ for a solution $x^* \in \mathbb{R}^n$ to the problem if $\mathcal{S}(x) = \mathcal{S}(x^*)$.

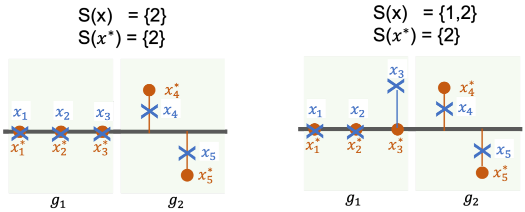


Figure: Support identification. The solution $x^* \in \mathbb{R}^5$ with group structures $g_1 = \{1, 2, 3\}$ and $g_2 = \{4, 5\}$. Support identification happens at the $x \in \mathbb{R}^5$ for the left figure while not for the right one.

Goal

Design an algorithm that can **simultaneously**

- achieve variance reduction
  - X full gradient evaluation
  - X storing a gradient table
- establish the support identification in the stochastic setting

Outline

Algorithm S-PStorm[3]

for $k = 1, 2, \ldots,$ **do**
  Draw m i.i.d samples $\{\xi_{k1}, \cdots, \xi_{km}\}$ w.r.t. $\mathcal{P}$.
  Set $v_k \leftarrow \frac{1}{m}\sum_{i=1}^{m} \nabla \ell(x_k; \xi_{ki})$.
  if $k = 1$ **then**
    Set $d_k \leftarrow v_k$.
  else
    $u_k \leftarrow \frac{1}{m}\sum_{i=1}^{m} \nabla \ell(x_{k-1}; \xi_{ki})$.
    Set $d_k \leftarrow v_k + (1 - \beta_k)(d_{k-1} - u_k)$.
  end if
  Compute $y_k \leftarrow \arg\min_{x \in \mathbb{R}^n} \left\{ \phi_p(x; x_k, \alpha_k, d_k) := \frac{1}{2\alpha_k} \|x - \left(x_k - \alpha_k d_k\right)\|_2^2 + r(x) \right\}$.
  Set $x_{k+1} \leftarrow x_k + \zeta\beta_k(y_k - x_k)$.
end for

[3] Yangyang Xu and Yibo Xu. "Momentum-based variance-reduced proximal stochastic gradient method for composite nonconvex stochastic optimization". In: *Journal of Optimization Theory and Applications* 196.1 (2023), pp. 266–297.

Algorithm `S-PStorm`[3]

for $k = 1, 2, \ldots,$ **do**
  Draw m i.i.d samples $\{\xi_{k1}, \cdots, \xi_{km}\}$ w.r.t. $\mathcal{P}$.
  Set $v_k \leftarrow \frac{1}{m}\sum_{i=1}^{m} \nabla \ell(x_k; \xi_{ki})$.
  if $k = 1$ **then**
    Set $d_k \leftarrow v_k$.
  else
    $u_k \leftarrow \frac{1}{m}\sum_{i=1}^{m} \nabla \ell(x_{k-1}; \xi_{ki})$.
    Set $d_k \leftarrow v_k + (1-\beta_k)(d_{k-1} - u_k)$.
  end if
  Compute $y_k \leftarrow \arg\min_{x\in\mathbb{R}^n} \left\{ \phi_p(x; x_k, \alpha_k, d_k) := \frac{1}{2\alpha_k}\|x - (x_k - \alpha_k d_k)\|_2^2 + r(x) \right\}$.
  Set $x_{k+1} \leftarrow x_k + \zeta\beta_k(y_k - x_k)$.
end for

Inexact Proximal Operator Evaluation?

$$y_k \approx_{\tilde{\varepsilon}_k} \arg\min_{x\in\mathbb{R}^n} \left\{ \frac{1}{2\alpha_k}\|x - (x_k - \alpha_k d_k)\|_2^2 + r(x) \right\}.$$

Definition of $\tilde{\varepsilon}_k$-inexact solution:

$$\phi_p(y_k; x_k, \alpha_k, d_k) \le \phi_p(y_k^*; x_k, \alpha_k, d_k) + \tilde{\varepsilon}_k \text{ where } y_k^* \text{ is the solution.}$$

[3] Yangyang Xu and Yibo Xu. "Momentum-based variance-reduced proximal stochastic gradient method for composite nonconvex stochastic optimization". In: *Journal of Optimization Theory and Applications* 196.1 (2023), pp. 266–297.

Outline

Key Lemma: bounding the error $\epsilon_k = d_k - \nabla f(x_k)$

Tail bound.

[Rephrased from[a].] Suppose $\{S_t\}_{t=0}^{\infty}$ forms a martingale and denote $e_t = S_t - S_{t-1}$. If $\sum_{t=1}^{\infty} \|e_t\|_{\infty}^2 \leq \texttt{const}$ almost surely. Then for $\rho > 0$,

$$\mathbb{P}\left[\sup_t \|S_t\| \geq \rho\right] \leq 2\exp\left(-\frac{\rho^2}{2\texttt{const}^2}\right).$$

[a] Iosif Pinelis. "Optimum bounds for the distributions of martingales in Banach spaces". In: *The Annals of Probability* (1994), pp. 1679–1706.

Key Lemma: bounding the error $\epsilon_k = d_k - \nabla f(x_k)$

Tail bound.

[Rephrased from[a].] Suppose $\{S_t\}_{t=0}^{\infty}$ forms a martingale and denote $e_t = S_t - S_{t-1}$. If $\sum_{t=1}^{\infty} \|e_t\|_{\infty}^2 \leq \mathtt{const}$ almost surely. Then for $\rho > 0$,

$$\mathbb{P}\left[\sup_t \|S_t\| \geq \rho\right] \leq 2\exp\left(-\frac{\rho^2}{2\mathtt{const}^2}\right).$$

[a] Iosif Pinelis. “Optimum bounds for the distributions of martingales in Banach spaces”. In: *The Annals of Probability* (1994), pp. 1679–1706.

- Decompose $\epsilon_k = d_k - \nabla f(x_k) = \sum_{t=0}^{k} e_{kt}$.
- Define $S_{kt} = \sum_{i=0}^{t} e_{ki}$ for all $0 \leq t \leq k$. Observe that $\epsilon_k = S_{kk}$.
- Derive the upper bound of $\sum_{t=1}^{k} \|e_{kt}\|^2$.

A high probability bound on ϵ_k.

Algorithmic Choices: $\beta_k = \min\{1/2, c/(k+1)\}$ with $c > 1$ and $\alpha_k \equiv \underline{\alpha}$for all $k \geq 1$.
Let $\eta_k > 0$, and define $\underline{k} = \lceil (2c) - 1 \rceil$ and

$$U(k) = \Theta\left(\max\left\{\left(\frac{\underline{k}+1}{k+2}\right)^c, \frac{c}{\sqrt{k+2}}\right\}\sqrt{\log\frac{2}{\eta_k}}\right)$$

Theorem 1

Under certain assumptions, let $\eta_k = \frac{\eta_0}{k^2}$ *for all* $k \geq 1$ *with* $\eta_0 \in (0, 6/\pi^2)$*, then*

$$\mathbb{P}\left[\bigcap_{k \geq \underline{k}}^{\infty} \{\|\epsilon_k\| \leq U(k)\}\right] \geq 1 - \frac{\eta_0 \pi^2}{6}.$$

$$U(k) = \Theta(\max\{\sqrt{\log k}/k^c, \sqrt{\log k / k}\})$$

Iterates Convergence

Additional Assumption: f is μ_f strongly convex.
Algorithmic choice: Let $\underline{\alpha} = \mu_f / L_g^2$, $\zeta \in (0, 2)$, $\theta \geq 2$, $c = (2\theta L_g^2)/(\zeta \mu_f^2) > 2$, and $\underline{k} = \lceil 2c - 1 \rceil$. Set $\eta_k = \eta_0 / k^2$ for all $k \geq 1$ with $\eta_0 \in (0, 6/\pi^2)$.

Theorem 2 (exact proximal operator evaluation)

$$\mathbb{P}\left[\bigcap_{k \geq \underline{k}}^{\infty} \left\{ \|x_k - x^*\|^2 \leq \bar{c}_1 \frac{\|x_{\underline{k}} - x^*\|^2}{k^\theta} + \bar{c}_2 \frac{\log \frac{2k}{\eta_0}}{k} \right\}\right] \geq 1 - \eta_0 \pi^2 / 6 > 0.$$

Iterates Convergence

Additional Assumption: f is μ_f strongly convex.
Algorithmic choice: Let $\underline{\alpha} = \mu_f / L_g^2$, $\zeta \in (0,2)$, $\theta \geq 2$, $c = (2\theta L_g^2)/(\zeta \mu_f^2) > 2$, and $\underline{k} = \lceil 2c - 1 \rceil$. Set $\eta_k = \eta_0 / k^2$ for all $k \geq 1$ with $\eta_0 \in (0, 6/\pi^2)$.

Theorem 2 (exact proximal operator evaluation)

$$\mathbb{P}\left[\bigcap_{k \geq \underline{k}}^{\infty} \left\{ \|x_k - x^*\|^2 \leq \bar{c}_1 \frac{\|x_{\underline{k}} - x^*\|^2}{k^\theta} + \bar{c}_2 \frac{\log \frac{2k}{\eta_0}}{k} \right\}\right] \geq 1 - \eta_0 \pi^2/6 > 0.$$

Theorem 3 (inexact proximal operator evaluation)

$$\mathbb{P}\left[\bigcap_{k \geq \underline{k}}^{\infty} \left\{ \|x_k - x^*\|^2 \leq \bar{c}_1' \frac{\|x_{\underline{k}} - x^*\|^2}{k^\theta} + \bar{c}_2' \frac{\log \frac{2k}{\eta_0}}{k+1} + \bar{c}_3' A_k \right\}\right] \geq 1 - \eta_0 \pi^2/6 > 0,$$

where $A_k := \frac{1}{(k+1)^\theta} \cdot \sum_{i=1}^{k} (i+3)^\theta \tilde{\varepsilon}_i$ *and* $\{\tilde{\varepsilon}_i\}$ *measure the inexactness of the proximal operator evaluation.*

Choose $\tilde{\varepsilon}_i = \log(i+1)/(i+1)^2$ for all i to recover the complexity for the exact case.

Definition: Support Identification in the Stochastic Setting

Support identification in stochastic setting can be defined in the **expectation sense**[4], in the **high-probability sense**[5], and in the **almost surely sense**[6].

Definition 4 (support identification with high probability)

There exist $K \in \mathbb{N}_+$ and $p \in (0, 1]$ such that

$$\mathbb{P}\left[\{\mathcal{S}(x_k) = \mathcal{S}(x^*)\}\right] \geq 1 - p \text{ for each } k \geq K.$$

[4] Yifan Sun et al. "Are we there yet? Manifold identification of gradient-related proximal methods". In: *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*. Ed. by Kamalika Chaudhuri and Masashi Sugiyama. Vol. 89. Proceedings of Machine Learning Research. PMLR, 2019, pp. 1110–1119.

[5] Sangkyun Lee and Stephen J Wright. "Manifold Identification in Dual Averaging for Regularized Stochastic Online Learning.". In: *Journal of Machine Learning Research* 13.6 (2012).

[6] Clarice Poon, Jingwei Liang, and Carola Schoenlieb. "Local convergence properties of SAGA/Prox-SVRG and acceleration". In: *International Conference on Machine Learning*. PMLR. 2018, pp. 4124–4132.

Definition: Support Identification in the Stochastic Setting

Support identification in stochastic setting can be defined in the **expectation sense**[4], in the **high-probability sense**[5], and in the **almost surely sense**[6].

Definition 4 (support identification with high probability)

There exist $K \in \mathbb{N}_+$ and $p \in (0, 1]$ such that

$$\mathbb{P}\left[\{\mathcal{S}(x_k) = \mathcal{S}(x^*)\}\right] \geq 1 - p \text{ for each } k \geq K.$$

Definition 5 (**consistent** identification with high probability)

There exist $K \in \mathbb{N}_+$ and $p \in (0, 1]$ such that

$$\mathbb{P}\left[\bigcap_{k \geq K}^{\infty} \{\mathcal{S}(x_k) = \mathcal{S}(x^*)\}\right] \geq 1 - p.$$

[4] Yifan Sun et al. "Are we there yet? Manifold identification of gradient-related proximal methods". In: *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*. Ed. by Kamalika Chaudhuri and Masashi Sugiyama. Vol. 89. Proceedings of Machine Learning Research. PMLR, 2019, pp. 1110–1119.

[5] Sangkyun Lee and Stephen J Wright. "Manifold Identification in Dual Averaging for Regularized Stochastic Online Learning.". In: *Journal of Machine Learning Research* 13.6 (2012).

[6] Clarice Poon, Jingwei Liang, and Carola Schoenlieb. "Local convergence properties of SAGA/Prox-SVRG and acceleration". In: *International Conference on Machine Learning*. PMLR. 2018, pp. 4124–4132.

Support Identification

Theorem 6

Under all previous assumptions and algorithmic choices, there exists constants $\{C_1, C_2, C_3\} \subseteq \mathbb{R}^n_+$ *that are independent of* k, $k_{\Delta^*} = \left(\frac{C_2}{\Delta^*}\right)^4$ *and* $k_{\delta^*} = \left(\frac{C_1}{\delta^*}\right)^{4/C_3}$ *such that, with* $K := \max\{k_{\Delta^*}, k_{\delta^*}, \underline{k}\}$*, it follows that*

$$\mathbb{P}\left[\bigcap_{k\geq K}^{\infty} \{\mathcal{S}(y_k) = \mathcal{S}(x^*)\}\right] \geq 1 - \frac{\eta_0 \pi^2}{6} > 0.$$

- $\Delta^* \in (0,1)$ measures the primal non-degeneracy;
- $\delta^* \in (0,1)$ measures the dual non-degeneracy;
- exact proximal operator evaluation ($\tilde{\varepsilon}_k = 0$ for all k): $C_3 = 1$;
- inexact proximal operator evaluation ($\tilde{\varepsilon}_k = \frac{\log k}{(k+3)^2}$ for all k): $0 < C_3 < 1$.

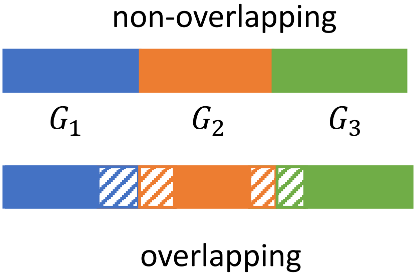
Summary

Algorithm	$\|x_k - x^*\|^2$	Support Identification	# Exact ∇f	Storage
ProxSVRG	$\mathcal{O}\left(\rho^k_{\texttt{ProxSVRG}}\right)$	$\mathcal{O}(\log(1/\delta^*))$	every epoch	$\mathcal{O}(n)$
SAGA	$\mathcal{O}\left(\rho^k_{\texttt{SAGA}}\right)$	$\mathcal{O}(\log(1/\delta^*))$	once	$\mathcal{O}(Nn)$
RDA	$\mathcal{O}(\log k/k)$	$\mathcal{O}\left(\frac{1}{(\delta^*)^4}\right)$	never	$\mathcal{O}(n)$
S-PStorm	$\mathcal{O}(\log k/k)$	$\mathcal{O}\left(\max\left\{\frac{1}{(\delta^*)^4}, \frac{1}{(\Delta^*)^4}\right\}\right)$	never	$\mathcal{O}(n)$

Table: Comparison of the complexity of different methods assuming the exact proximal operator evaluation.

Outline

Different Group Structures



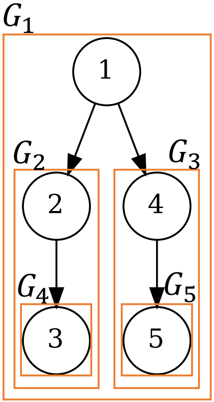


Figure: Left: Chain-like group structure; Right: Tree-like group structure

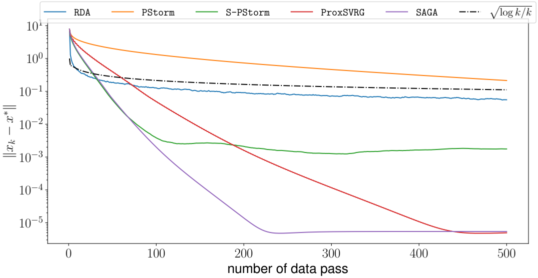
Problem

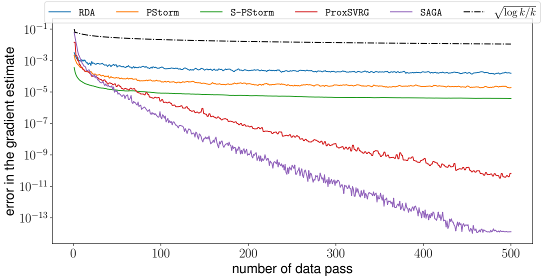
$$\min_{x\in\mathbb{R}^n} \frac{1}{N}\sum_{j=1}^{N} \log\left(1+e^{-y_j x^T d_j}\right) + 10^{-5}\|x\|^2 + \sum_{i=1}^{n_{\mathcal{G}}} \lambda_i \left\|[x]_{g_i}\right\|$$

data set	N	n
a9a	32561	123
avazu-app.tr	12,642,186	1,000,000
covtype	581,012	54
kdd2010	8,407,752	20,216,830
news20	19,996	1,355,191
phishing	11,055	68
rcv1	20,242	47,236
real-sim	72,309	20,958
url	2,396,130	3,231,961
w8a	49,749	300

- N is the number of data points, $d_j \in \mathbb{R}^n$ is the jth data point, $y_j \in \{-1, 1\}$ is the class label
- non-overlapping chain structure (more in the paper)
- $n_{\mathcal{G}} \in \{\lfloor 0.25n \rfloor, \lfloor 0.50n \rfloor, \lfloor 0.75n \rfloor, n\}$.
- $\Lambda = 0.1\Lambda_{\text{min}}$ and $\Lambda = 0.01\Lambda_{\text{min}}$.

Iterates and Error Convergence





Support Identification

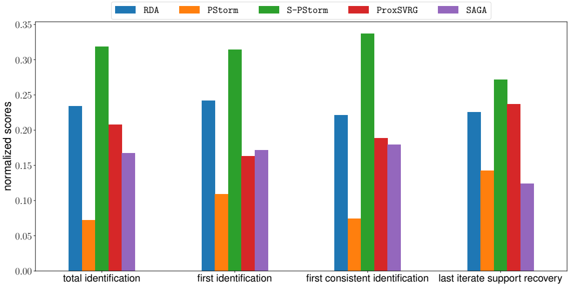


Figure: Normalized scores for four metrics that evaluate the performance of the support identification.

Summary

We designed an algorithm that can **simultaneously** achieve

- variance reduction without any full gradient evaluation and storing a huge gradient table,
- **consistent** support identification, and
- strong empirical performance.

Q&A

Thank you and Questions?

Contact: yud319@lehigh.edu

A high probability bound on ϵ_k.

Algorithmic Choices: $\beta_k = \min\{1/2, c/(k+1)\}$ with $c > 1$ and $\alpha_k \equiv \underline{\alpha}$ for all $k \geq 1$.
Let $\eta_k > 0$, and define $\underline{k} = \lceil (2c) - 1 \rceil$ and

$$U(k) = \Theta\left(\max\left\{\left(\frac{\underline{k}+1}{k+2}\right)^c, \frac{c}{\sqrt{k+2}}\right\}\sqrt{\log\frac{2}{\eta_k}}\right)$$

Theorem 7

Under certain assumptions, let $\eta_k = \frac{\eta_0}{k^2}$ *for all* $k \geq 1$ *with* $\eta_0 \in (0, 6/\pi^2)$*, then*

$$\mathbb{P}\left[\bigcap_{k \geq \underline{k}}^{\infty} \{\|\epsilon_k\| \leq U(k)\}\right] \geq 1 - \frac{\eta_0 \pi^2}{6}.$$

- ∇f is L_g-Lipschitz continuous and r is convex and closed
- $\mathbb{E}_{\xi \sim \mathcal{P}}[\nabla \ell(x_k; \xi) \mid \mathcal{F}_k] = \nabla f(x_k)$
- $\mathbb{P}_{\xi \sim \mathcal{P}}\{\|\nabla \ell(x_k, \xi) - \nabla f(x_k)\| \leq \sigma \mid \mathcal{F}_k\} = 1$
- $\mathbb{P}_{\xi \sim \mathcal{P}}\{\|d_k\| \leq G_d \mid \mathcal{F}_k\} = 1$
- $\mathbb{P}\{\|g_r\|_2 \leq G_r, \ \forall g_r \in \partial r(x_k)\} = 1$

Error Decomposition

For all $k \geq 2$, with the convention that $\prod_{i=l}^{u} a_i = 1$ if $l > u$, consider $\{e_{ki}\}_{i=0}^{k}$ with

$$e_{ki} := \begin{cases} 0 & i = 0, \\ \left(\prod_{j=2}^{k}(1-\beta_j)\right) A_1 & i = 1, \\ \left(\prod_{j=i+1}^{k}(1-\beta_j)\right) A_i + \left(\prod_{j=i}^{k}(1-\beta_j)\right) B_i & 2 \leq i \leq k, \end{cases}$$

where $A_i := v_i - \nabla f(x_i)$ and $B_i := \nabla f(x_{i-1}) - u_i$ for all $i \geq 1$ with v_i and u_i defined as in Algorithm 1.