Stochastic Approximation Methods for Nonconvex Stochastic Composite Optimization

Hongchao Zhang
hozhang@math.lsu.edu

Department of Mathematics
Center for Computation and Technology
Louisiana State University

Joint work with S. Ghadimi & G. Lan

December 22nd, 2013

Stochastic Composite Optimization

Optimize

$$\min_{\mathbf{x}\in\mathbf{X}} F(\mathbf{x}) := f(\mathbf{x}) + \phi(\mathbf{x}),$$

where

- $f \in \mathcal{C}_L^{1,1}(\mathbf{X})$, but ∇f is not available. $\mathbf{X} \in \mathbb{R}^n$ is a convex set.
- For any $\mathbf{x}_k \in \mathbf{X}$, a *stochastic first-order oracle* ($\mathcal{SFO}$) provides a *stochastic gradient* $G(\mathbf{x}_k, \xi_k)$, or a *stochastic zero-order oracle* ($\mathcal{SZO}$) provides a *stochastic function value* $F(\mathbf{x}_k, \xi_k)$, where ξ_k is a random variable supported on Ξ_k.
- ϕ is a simple convex function, but possibly nonsmooth. (Ex. $\phi = \|\cdot\|_1, \phi = \|\cdot\|_{TV}$ or $\phi \equiv 0$.)

Outline

- The Generalized Projection and its Properties
- The Stochastic First-order methods
  (Stochastic Projected Gradient Method)
- The Stochastic Zero-order methods
  (Stochastic Projected Gradient-free Method)
- Preliminary Numerical Results

The generalized projection

- The (generalized) projection:

$$\mathbf{x}^+(\mathbf{x}, \mathbf{g}, \gamma) = \operatorname{Arg}\min_{\mathbf{u}\in\mathbf{X}} \left\{ \langle \mathbf{g}, \mathbf{u} \rangle + \frac{1}{\gamma} V(\mathbf{u}, \mathbf{x}) + \phi(\mathbf{u}) \right\},$$

where $\gamma > 0$, V is the *prox-function* associated with $\omega \in \mathcal{S}^{1,1}_{\nu,L}$

$$V(\mathbf{u}, \mathbf{x}) := \omega(\mathbf{u}) - [\omega(\mathbf{x}) + \langle \nabla\omega(\mathbf{x}), \mathbf{u} - \mathbf{x} \rangle].$$

Ext. $\omega(\mathbf{x}) = \|\mathbf{x}\|^2/2$ with $\nu = 1$, then $V(\mathbf{u}, \mathbf{x}) = \|\mathbf{u} - \mathbf{x}\|^2/2$.

- Assumption: The (generalized) projection is relatively easily solvable.

Properties of the projection

- Definition: Let $P_{\mathbf{X}}(\mathbf{x}, \mathbf{g}, \gamma) = \frac{1}{\gamma}(\mathbf{x} - \mathbf{x}^+)$.

- For any $\mathbf{x} \in \mathbf{X}$, $\mathbf{g} \in \mathbb{R}^n$ and $\gamma > 0$, we have

$$\langle \mathbf{g}, P_{\mathbf{X}}(\mathbf{x}, \mathbf{g}, \gamma) \rangle \geq \nu \|P_{\mathbf{X}}(\mathbf{x}, \mathbf{g}, \gamma)\|^2 + \frac{1}{\gamma}\left[h(\mathbf{x}^+) - h(\mathbf{x})\right].$$

- If $\mathbf{x}_1^+ = \mathbf{x}^+(\mathbf{x}, \mathbf{g}_1, \gamma)$ and $\mathbf{x}_2^+ = \mathbf{x}^+(\mathbf{x}, \mathbf{g}_2, \gamma)$, then

$$\|\mathbf{x}_2^+ - \mathbf{x}_1^+\| \leq \frac{\gamma}{\nu}\|\mathbf{g}_2 - \mathbf{g}_1\|$$

  and

$$\|P_{\mathbf{X}}(\mathbf{x}, \mathbf{g}_1, \gamma) - P_{\mathbf{X}}(\mathbf{x}, \mathbf{g}_2, \gamma)\| \leq \frac{1}{\nu}\|\mathbf{g}_1 - \mathbf{g}_2\|.$$

Properties of the projection

- For any $\mathbf{u} \in \mathbf{X}$, we have

$$\begin{aligned} & \langle \mathbf{g}, \mathbf{x}^+ \rangle + h(\mathbf{x}^+) + \frac{1}{\gamma} V(\mathbf{x}^+, \mathbf{x}) \\ \leq \; & \langle \mathbf{g}, \mathbf{u} \rangle + h(\mathbf{u}) + \frac{1}{\gamma} [V(\mathbf{u}, \mathbf{x}) - V(\mathbf{u}, \mathbf{x}^+)]. \end{aligned}$$

The Stochastic First-order methods

Assumption:

- For any $k \geq 1$, we have

$$\begin{aligned} &\text{a)} \quad \mathbb{E}[G(\mathbf{x}_k, \xi_k)] = \nabla f(\mathbf{x}_k) \\ &\text{b)} \quad \mathbb{E}\left[\|G(\mathbf{x}_k, \xi_k) - \nabla f(\mathbf{x}_k)\|^2\right] \leq \sigma^2, \end{aligned}$$

for some $\sigma > 0$.

A general RSPG Algorithm

Input: Initial point $\mathbf{x}_1 \in \mathbf{X}$, iteration limit N, the stepsizes $\{\gamma_k > 0\}$, the batch sizes $\{m_k\}$, and the probability mass function P_R supported on $\{1, \ldots, N\}$.

Step 0. Let R be a random variable with density function P_R.

Step $k = 1, \ldots, R-1$. Call the $\mathcal{SFO}$ m_k times to obtain $G(\mathbf{x}_k, \xi_{k,i})$, $i = 1, \ldots, m_k$, and set $G_k = (\sum_{i=1}^{m_k} G(\mathbf{x}_k, \xi_{k,i}))/m_k$, and compute

$$\mathbf{x}_{k+1} = \operatorname{Arg}\min_{\mathbf{u}\in\mathbf{X}} \left\{ \langle G_k, \mathbf{u}\rangle + \frac{1}{\gamma_k} V(\mathbf{u}, \mathbf{x}_k) + \phi(\mathbf{u}) \right\}.$$

Output: $\mathbf{x}_R$.

Convergence Complexity

Theorem. Suppose

- $\{\gamma_k\}$ satisfy $0 < \gamma_k \le \nu/L$, $\gamma_k < \nu/L$ for at least one k,
- $P_R(k) = t_k / \sum_{k=1}^N t_k$, where $t_k = \nu\gamma_k - L\gamma_k^2$.

Then, we have

$$\mathbb{E}[\|\tilde{\mathbf{g}}_{\mathbf{x},R}\|^2] \le \left[L D_F^2 + \frac{\sigma^2}{\nu} \sum_{k=1}^N (\gamma_k / m_k) \right] / \sum_{k=1}^N t_k,$$

where the expectation is w.r.t. R and $\xi_{[N]} := (\xi_1, \ldots, \xi_N)$,
$D_F = \sqrt{(F(\mathbf{x}_1) - F^*)/L}$ and $\tilde{\mathbf{g}}_{\mathbf{x},R} = P_{\mathbf{X}}(\mathbf{x}_R, G_R, \gamma_R)$.
In addition, if f is convex and $0 < \gamma_k \le \ldots \le \gamma_N \le \nu/L$, then

$$\mathbb{E}[F(\mathbf{x}_R) - F^*] \le \left((\nu - L\gamma_1) V(\mathbf{x}^*, \mathbf{x}_1) + \frac{\sigma^2}{2} \sum_{k=1}^N \frac{\gamma_k^2}{m_k} \right) / \sum_{k=1}^N t_k.$$

Convergence Complexity

Comment:

- If f is convex, the batch size $m_k = 1$, by choosing $\gamma_k = \mathcal{O}(1/\sqrt{k})$ we still get sub-optimal convergence rate $\mathbb{E}[F(\mathbf{x}_R) - F^*] \leq \mathcal{O}(\ln N/\sqrt{N})$.
- If f is nonconvex and $m_k = 1$, regardless of choice γ_k, we can not guarantee convergence.
- If we choose $\gamma_k = \nu/L$ and $m_k = m$, we have

$$\mathbb{E}[\|\tilde{g}_{\mathbf{x},R}\|^2] \leq \frac{4L^2 D_F^2}{\nu^2 N} + \frac{2\sigma^2}{\nu^2 m}$$

and if f is convex, we have

$$\mathbb{E}[f(\mathbf{x}_R) - f^*] \leq \frac{2LV(\mathbf{x}^*, \mathbf{x}_1)}{N\nu} + \frac{\sigma^2}{2Lm}$$

Convergence Complexity

Corollary. Given total budget $\bar{N}$ calls of $\mathcal{SFO}$. Suppose $\gamma_k = \nu/(2L)$ and $m_k = m := \min\{\lceil \max\{1, \sigma\sqrt{6\bar{N}}/(4L\tilde{D})\}\rceil, \bar{N}\}$ with $\bar{N} \geq 3\sigma^2/(8L^2\tilde{D}^2)$. Then, if $\tilde{D} = D_F$, we have

$$(\nu^2/L)\mathbb{E}[\|\mathbf{g}_{\mathbf{x},R}\|^2] \leq \mathcal{B}_{\bar{N}} := \frac{16L^2D_F^2}{\bar{N}} + \frac{8\sqrt{6}D_F\sigma}{\sqrt{\bar{N}}}.$$

If f is convex and $\tilde{D} = \sqrt{3V(\mathbf{x}^*, \mathbf{x}_1)/\nu}$, then

$$\mathbb{E}[F(\mathbf{x}_R) - F^*] \leq \frac{4LV(\mathbf{x}^*, \mathbf{x}_1)}{\nu\bar{N}} + \frac{2\sqrt{2V(\mathbf{x}^*, \mathbf{x}_1)}\sigma}{\sqrt{\nu\bar{N}}}.$$

Comment:

- Optimal ! The second term is unimprovable. (Nemirovski, 1983)

A two-phase stochastic projected gradient algorithm

- Definition: An (ϵ, Λ)-solution: $\mathbf{x} \in \mathbf{X}$ such that

$$\text{Prob}\{[\|\mathbf{g_x}(\mathbf{x})\|^2 \leq \epsilon\} \geq 1 - \Lambda,$$

where $\epsilon > 0$, $\Lambda \in (0, 1)$ and $\mathbf{g_x}(\mathbf{x}) = P_{\mathbf{X}}(\mathbf{x}, \nabla f(\mathbf{x}), \gamma)$.

- Let $\gamma_k = \gamma := \nu/(2L)$ and $m_k = m$, by Markov's inequality

$$\text{Prob}\left\{\|\mathbf{g}_{\mathbf{x},R}\|^2 \geq \frac{\lambda L \mathcal{B}_{\bar{N}}}{\nu^2}\right\} \leq \frac{1}{\lambda}, \qquad \text{for any } \lambda > 0.$$

- An (ϵ, Λ)-solution can be bounded by

$$\mathcal{O}\left\{\frac{1}{\Lambda\epsilon} + \frac{\sigma^2}{\Lambda^2\epsilon^2}\right\}.$$

A two-phase RSPG Algorithm

Input: Initial point $\mathbf{x}_1 \in \mathbf{X}$, number of runs S, total $\bar{N}$ of calls to the $\mathcal{SFO}$ in each run of the RSPG algorithm, and sample size T in the post-optimization phase.

Optimization phase: For $s = 1, \ldots, S$, call the RSPG algorithm with initial point x_1, iteration limit $N = \lfloor \bar{N}/m \rfloor$ and $\gamma_k = \nu/(2L)$.

Post-optimization phase: Choose a solution $\bar{\mathbf{x}}^*$ from the candidate list $\{\bar{\mathbf{x}}_1, \ldots, \bar{\mathbf{x}}_S\}$ such that

$$\|\bar{\mathbf{g}}_{\mathbf{X}}(\bar{\mathbf{x}}^*)\| = \min_{s=1,\ldots,S} \|\bar{\mathbf{g}}_{\mathbf{X}}(\bar{\mathbf{x}}_s)\|, \quad \bar{\mathbf{g}}_{\mathbf{X}}(\bar{\mathbf{x}}_s) := P_{\mathbf{X}}(\bar{\mathbf{x}}_s, \bar{G}_T(\bar{\mathbf{x}}_s), \gamma_{R_s}),$$

where $\bar{G}_T(\mathbf{x}) = \frac{1}{T}\sum_{k=1}^{T} G(\mathbf{x}, \xi_k)$.
Output: $\mathbf{x}_R$.

Convergence Complexity

Theorem. The following statements holds for 2-RSPG algorithm:

(a) For all $\lambda > 0$, we have

$$\text{Prob}\left\{\|\mathbf{g}_{\mathbf{x}}(\bar{\mathbf{x}}^*)\|^2 \geq \frac{2}{\nu^2}\left(4L\mathcal{B}_{\bar{N}} + \frac{3\lambda\sigma^2}{T}\right)\right\} \leq \frac{S}{\lambda} + 2^{-S};$$

(b) With a particular choice of $(S(\Lambda), T(\epsilon, \Lambda), \bar{N}(\epsilon))$, 2-RSPG finds an (ϵ, Λ)-solution with the number of calls of $\mathcal{SFO}$:

$$\mathcal{O}\left\{\frac{1}{\epsilon}\log_2\frac{1}{\Lambda} + \frac{\sigma^2}{\epsilon^2}\log_2\frac{1}{\Lambda} + \frac{\sigma^2}{\Lambda\epsilon}\log_2^2\frac{1}{\Lambda}\right\}.$$

Comment:

- The second term smaller to a factor of $1/[\Lambda^2\log_2(1/\Lambda)]$.

Convergence Complexity

Under a "Light-tail" assumption: for any $\mathbf{x}_k \in \mathbf{X}$, we have

$$\mathbb{E}[\exp\{\|G(\mathbf{x}_k, \xi_k) - \nabla f(\mathbf{x}_k)\|^2/\sigma^2\}] \le \exp\{1\},$$

(a) for all $\lambda > 0$, we have

$$\text{Prob}\left\{\|\mathbf{g}_{\mathbf{X}}(\bar{\mathbf{x}}^*)\|^2 \ge \left[\frac{8L\mathcal{B}_{\bar{N}}}{\nu^2} + \frac{12(1+\lambda)^2\sigma^2}{T\nu^2}\right]\right\} \le S\exp(-\frac{\lambda^2}{3}) + 2^{-S}$$

(b) With a particular choice of $(S(\Lambda), T(\epsilon, \Lambda), \bar{N}(\epsilon))$, 2-RSPG finds an (ϵ, Λ)-solution with the number of calls of $\mathcal{SFO}$:

$$\mathcal{O}\left\{\frac{1}{\epsilon}\log_2\frac{1}{\Lambda} + \frac{\sigma^2}{\epsilon^2}\log_2\frac{1}{\Lambda} + \frac{\sigma^2}{\epsilon}\log_2^2\frac{1}{\Lambda}\right\}.$$

Comment:

- The third term smaller to a factor of $1/\Lambda$.

The Stochastic Zero-order methods

- Assumption: For any $k \geq 1$, we have

$$\mathbb{E}[F(\mathbf{x}_k, \xi_k)] = f(\mathbf{x}_k) \text{ and } F(\cdot, \xi_k) \in \mathcal{C}_L^{1,1}(\mathbb{R}^n) \text{ almost surely.}$$

- Definition: A smooth Gaussian approximation of f

$$f_\mu(\mathbf{x}) := \frac{1}{(2\pi)^{\frac{n}{2}}} \int f(\mathbf{x} + \mu\mathbf{v}) e^{-\frac{1}{2}\|\mathbf{v}\|^2} \, d\mathbf{v} = \mathbb{E}_\mathbf{v}[f(\mathbf{x} + \mu\mathbf{v})],$$

  where $\mathbf{v}$ is a n-dimensional standard Gaussian random vector.

- Definition: the approximated stochastic gradient of f at $\mathbf{x}_k$

$$G_\mu(\mathbf{x}_k, \xi_k, \mathbf{v}) := \frac{F(\mathbf{x}_k + \mu\mathbf{v}, \xi_k) - F(\mathbf{x}_k, \xi_k)}{\mu} \mathbf{v}.$$

Comment: Nesterov, 2010.
$f_\mu \in \mathcal{C}_{L_\mu}^{1,1}(\mathbb{R}^n)$ with $L_\mu \leq L$ and $\mathbb{E}_{\mathbf{v},\xi_k}[G_\mu(\mathbf{x}_k, \xi_k, \mathbf{v})] = \nabla f_\mu(\mathbf{x}_k)$.

A randomized stochastic gradient free algorithm

A general RSGF Algorithm

Input: Initial point $\mathbf{x}_1 \in \mathbf{X}$, iteration limit N, the stepsizes $\{\gamma_k > 0\}$, the batch sizes $\{m_k\}$, and the probability mass function P_R supported on $\{1, \ldots, N\}$.

Step 0. Let R be a random variable with density function P_R.

Step $k = 1, \ldots, R-1$. Call the $\mathcal{SZO}$ m_k times to obtain $G_{\mu,k} = (\sum_{i=1}^{m_k} G_\mu(\mathbf{x}_k, \xi_{k,i}, \mathbf{v}_{k,i}))/m_k$, and compute

$$\mathbf{x}_{k+1} = \operatorname{Arg}\min_{\mathbf{u} \in \mathbf{X}} \left\{ \langle G_{\mu,k}, \mathbf{u} \rangle + \frac{1}{\gamma_k} V(\mathbf{u}, \mathbf{x}_k) + \phi(\mathbf{u}) \right\}.$$

Output: $\mathbf{x}_R$.

Convergence Complexity

Thm. Given total budget $\bar{N}$ calls of $\mathcal{SZO}$. Suppose $\gamma_k = \nu/(2L)$ and $m_k = \min\{\lceil \max\{\sqrt{(n+4)(M^2+\sigma^2)\bar{N}/(L\tilde{D})}, n+4\}\rceil, \bar{N}\}$ with $\bar{N} \geq \max\{(n+4)^2(M^2+\sigma^2)/(L\tilde{D})^2, n+4\}$.

If $\mu \leq D_F/\sqrt{(n+4)\bar{N}}$ and $\tilde{D} = D_F$,then

$$(\nu^2/L)\mathbb{E}[\|\mathbf{g}_{\mathbf{x},R}\|^2] \leq \frac{65L^2D_F^2(n+4)}{\bar{N}} + \frac{64\sqrt{(n+4)(M^2+\sigma^2)}}{\sqrt{\bar{N}}}.$$

If f convex, $\mu \leq \sqrt{V(\mathbf{x}^*, \mathbf{x}_1)/(\nu(n+4)\bar{N})}$, ,$\tilde{D} = 2\sqrt{V(\mathbf{x}^*, \mathbf{x}_1)/\nu}$,

$$\mathbb{E}[F(\mathbf{x}_R) - F^*] \leq \frac{6LV(\mathbf{x}^*, \mathbf{x}_1)(n+4)}{\nu\bar{N}} + \frac{4\sqrt{V(\mathbf{x}^*, \mathbf{x}_1)(n+4)(M^2+\sigma^2)}}{\sqrt{\nu\bar{N}}}.$$

Comment:

- Number of calls of $\mathcal{SZO}$ to find $\mathbb{E}[F(\mathbf{x}_R) - F^*] \leq \epsilon$ is bounded by $\mathcal{O}(n/\epsilon^2)$, when ϵ sufficiently small, better than $\mathcal{O}(n^2/\epsilon^2)$ by Nesterov, 2010.

Preliminary Numerical Results

- Algorithm schemes: Let $V(\mathbf{x},\mathbf{z}) = \|\mathbf{x}-\mathbf{z}\|^2/2$, $\gamma_k = 1/(2L)$. In 2-RSPG , we take $S = 5$ independent runs of RSPG and take $T = N/2$ in the post-optimization phase to choose the best $\bar{\mathbf{x}}^*$. The quality of $\bar{\mathbf{x}}^*$ is evaluated by i.i.d. sample of size $K >> \bar{N}$, where $\bar{N}$ is the iteration number in each RSPG.

- Estimation of parameters: Use i.i.d. sample of size $N_0 = 200$ to estimate L and σ. Since $F^* \geq 0$ in our example, we set $D_F = \sqrt{2F(\mathbf{x}_1)/L}$.

- Notations: NS is the maximum number of calls of stochastic oracle. Hence, $\bar{N} = NS$ in RSPG, and $\bar{N} = NS/S$ in 2-RSPG. $\bar{\mathbf{x}}^*$ is the output. *Mean* and *Var.* are the average and variants of the results over 20 runs of each algorithm.

Preliminary Numerical Results

- A least square problem with a smoothly clipped absolute deviation penalty term (Fan & Li, 2001):

$$\min_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x}) = \mathbb{E}_{\mathbf{u},\mathbf{v}}[(\langle \mathbf{x}, \mathbf{u}\rangle - \mathbf{v})^2] + \sum_{j=1}^{d} q_\lambda(|\mathbf{x}_j|),$$

  where $\mathbf{u}$ is drawn from standard normal, $v = \langle \bar{\mathbf{x}}, \mathbf{u}\rangle + \xi$ with $\xi \sim N(0, \bar{\sigma}^2)$ and $q_\lambda : \mathbb{R}_+ \to \mathbb{R}$, satisfying $q_\lambda(0) = 0$ with derivative defined as

$$q'_\lambda(\beta) = \left\{ \beta I(\beta \leq \lambda) + \frac{\max(0, a\lambda - \beta)}{(a-1)} I(\beta > \lambda) \right\}.$$

  Here $a > 2$ and $\lambda > 0$ are constant parameters.

- In numerical experiment, we set $a = 3.7$ and $\lambda = 0.1$, three different problem sizes with $n = 100, 500, 1000$ and two different noise levels with $\bar{\sigma} = 0.1, 1$.

Preliminary Numerical Results

Table: Estimated $\|\nabla f(\bar{\mathbf{x}}^*)\|^2$ for the least square problem ($K = 75,000$)

$\bar{N}S$		RSG	2-RSG	RSPG	2-RSPG
		$n = 100, \tilde{\sigma} = 0.1$			
1000	mean	0.2509	0.3184	0.1564	0.3176
	var.	4.31e-2	1.68e-2	4.58e-2	2.54e-2
5000	mean	0.0828	0.0841	0.0113	0.0164
	var.	6.75e-3	1.03e-3	4.22e-4	3.37e-4
25000	mean	0.0056	0.0070	0.0006	0.0010
	var.	1.69e-4	1.08e-4	2.05e-7	1.43e-7
		$n = 100, \tilde{\sigma} = 1$			
1000	mean	0.3731	0.3761	0.2379	0.3567
	var.	3.38e-2	1.40e-2	4.01e-2	1.41e-2
5000	mean	0.1095	0.1314	0.0436	0.0323
	var.	2.22e-2	3.96e-3	1.44e-2	8.69e-4
25000	mean	0.0374	0.0172	0.0138	0.0048
	var.	8.46e-3	1.83e-4	1.95e-3	8.48e-7

Preliminary Numerical Results

Table: Estimated $\|\nabla f(\bar{\mathbf{x}}^*)\|^2$ for the least square problem ($K = 75,000$)

$\bar{N}S$		RSG	2-RSG	RSPG	2-RSPG
		$n = 500, \tilde{\sigma} = 0.1$			
1000	mean	0.5479	0.6865	0.4212	0.8977
	var.	3.47e-2	6.17e-3	5.13e-2	2.64e-3
5000	mean	0.2481	0.3560	0.1030	0.1997
	var.	4.38e-2	3.45e-3	2.57e-2	2.21e-3
25000	mean	0.2153	0.0876	0.1093	0.0136
	var.	6.77e-2	1.13e-3	4.07e-2	3.24e-5
		$n = 500, \tilde{\sigma} = 1$			
1000	mean	0.5869	0.7444	0.4371	0.7771
	var.	2.14e-2	4.18e-3	3.40e-2	5.15e-3
5000	mean	0.3603	0.4732	0.1745	0.2987
	var.	3.77e-2	8.13e-3	3.51e-2	1.87e-2
25000	mean	0.2467	0.1584	0.1271	0.0351
	var.	6.49e-2	1.87e-3	4.30e-2	2.83e-4

Preliminary Numerical Results

Table: Estimated $\|\nabla f(\bar{\mathbf{x}}^*)\|^2$ for the least square problem $(K = 75,000)$

$\bar{N}S$		RSG	2-RSG	RSPG	2-RSPG
		$n = 1000, \tilde{\sigma} = 0.1$			
1000	mean	1.853	2.417	1.855	3.092
	var.	1.73e-1	1.31e-2	1.88e-1	1.29e-1
5000	mean	0.9555	1.501	0.4944	1.832
	var.	3.62e-1	6.39e-2	4.82e-1	2.36e-1
25000	mean	0.6305	0.4725	0.3402	0.1100
	var.	6.38e-1	2.08e-2	4.40e-1	4.54e-3
		$n = 1000, \tilde{\sigma} = 1$			
1000	mean	1.868	2.407	1.701	3.208
	var.	1.44e-1	1.22e-2	1.84e-1	1.54e-1
5000	mean	1.297	1.596	0.8032	1.403
	var.	5.25e-1	5.26e-2	6.38e-1	1.10e-1
25000	mean	0.575	0.6309	0.2079	0.1806
	var.	3.43e-1	4.65e-2	1.17e-1	1.43e-2

Preliminary Numerical Results

- A linear semi-supervised SVM problem (Chapelle et., 2008):

$$\min_{(\mathbf{x},b)\in\mathbb{R}^{n+1}} f(\mathbf{x},b) = \mathbb{E}_{\mathbf{u}_1,\mathbf{u}_2,v}[\lambda_1 \max\{0, 1 - v(\langle \mathbf{x}, \mathbf{u}_1\rangle + b)\}^2 + \lambda_2 e^{-5\{\langle \mathbf{x},\mathbf{u}_2\rangle + b\}^2}] + \lambda_3 \|\mathbf{x}\|_2^2.$$

  where $|b - 2r + 1| \leq \delta$, $\mathbf{u}_1$ and $\mathbf{u}_2$ are standard normal, $v \in \{0, 1\}$ with $v = \mathrm{sgn}(\langle \bar{\mathbf{x}}, \mathbf{u}_1\rangle + b)$ for some $\bar{\mathbf{x}} \in \mathbb{R}^n$. Here, λ_1, λ_2 and λ_3 are constant parameters, $r \in (0, 1)$ is the ration of positive labels and $\delta \in (0, 1)$ is the tolerance.
- In numerical experiment, we set $\lambda_1 = 1$, $\lambda_2 = \lambda_3 = 0.5$, $\delta = 0.1$ and three different problem sizes $n = 100, 500, 1000$.

Preliminary Numerical Results

Table: Estimated $\|\mathbf{g}_x(\bar{\mathbf{x}}^*)\|^2$ ($K = 75,000$)

$\bar{N}S$		RSPG	2-RSPG	RSPG	2-RSPG
		$n = 100$		$n = 500$	
1000	mean	1.355	0.2107	5.976	0.7955
	var.	1.21e+1	9.50e-3	1.93e+2	6.07e-1
5000	mean	0.1032	0.1174	0.2237	0.1703
	var.	4.96e-2	4.42e-3	1.93e+2	6.07e-1
25000	mean	0.0352	0.0699	0.2174	0.0832
	var.	1.13e-3	3.42e-3	2.35e-1	2.41e-4
		$n = 1000$			
1000	mean	27.06	2.417		
	var.	6.00e+3	1.73e+1		
5000	mean	16.24	0.4726		
	var.	2.20e+3	2.85e+1		
25000	mean	0.1007	0.1378		
	var.	2.46e-2	5.63e-5		