Approximate Proximal-Gradient Methods

Anis Hamadouche, Yun Wu, Andrew M. Wallace,

and Joaõ F. C. Mota

The Institute of Sensors, Signals & Systems (ISSS)

Heriot-Watt University, Edinburgh, UK

Composite Optimization Problems

Many problems in science, engineering, and *defense* can be written as

$$\underset{x \in \mathbb{R}^n}{\text{minimize}} \quad g(x) + h(x)$$

non-differentiable, can *encode constraints*: $h : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$

differentiable, real-valued: $g : \mathbb{R}^n \to \mathbb{R}$

Example: State-space model of a drone

$$x[k] = (\text{position time } k,\ \text{velocity time } k,\ \%\ \text{mission completed},\ \ldots)$$

$$x[k+1] \simeq Ax[k] + Bu[k]$$

input at time k

Goal: given $x[0]$, drive state to x_f in T time steps, while minimizing energy

$$\begin{array}{ll} \underset{\substack{x[1],\ldots,x[T] \\ u[0],\ldots,u[T-1]}}{\text{minimize}} & \left(x[T] - x_f\right)^2 + \sum_{k=0}^{T-1} u^2[k] \quad = g(x) \\ \text{subject to} & x[k+1] = Ax[k] + Bu[k]\,, \quad k = 0, \ldots, T-1 \end{array}$$

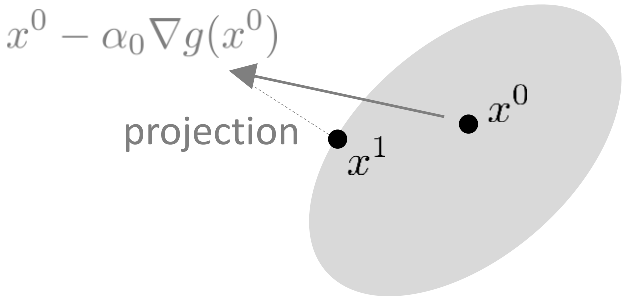
can be encoded in $h(x)$

Approximate Proximal Methods

$$\boxed{\underset{x \in \mathbb{R}^n}{\text{minimize}} \;\; f(x) := g(x) + h(x)}$$

$$x^{k+1} = \text{prox}_{\alpha_k h}\Big(\underbrace{x^k - \alpha_k \nabla g(x^k)}\Big)$$

gradient descent along g (differentiable)



$$\text{prox}_f(x) := \arg\min_y \; f(x) + \frac{1}{2}\|y - x\|_2^2$$

generalization of a *projection*

Errors may be introduced to *save power*

hardware, software, linear algebra, or algorithmic approximations

how to model ?

$$x^{k+1} = \text{prox}_{\alpha_k h}\Big(x^k - \alpha_k \nabla g(x^k) + \boldsymbol{\epsilon_1^k}\Big) + \boldsymbol{\epsilon_2^k}$$

Existing convergence proofs hold for this type of errors ?

Tradeoffs between power savings / accuracy / execution time ?

Related Work

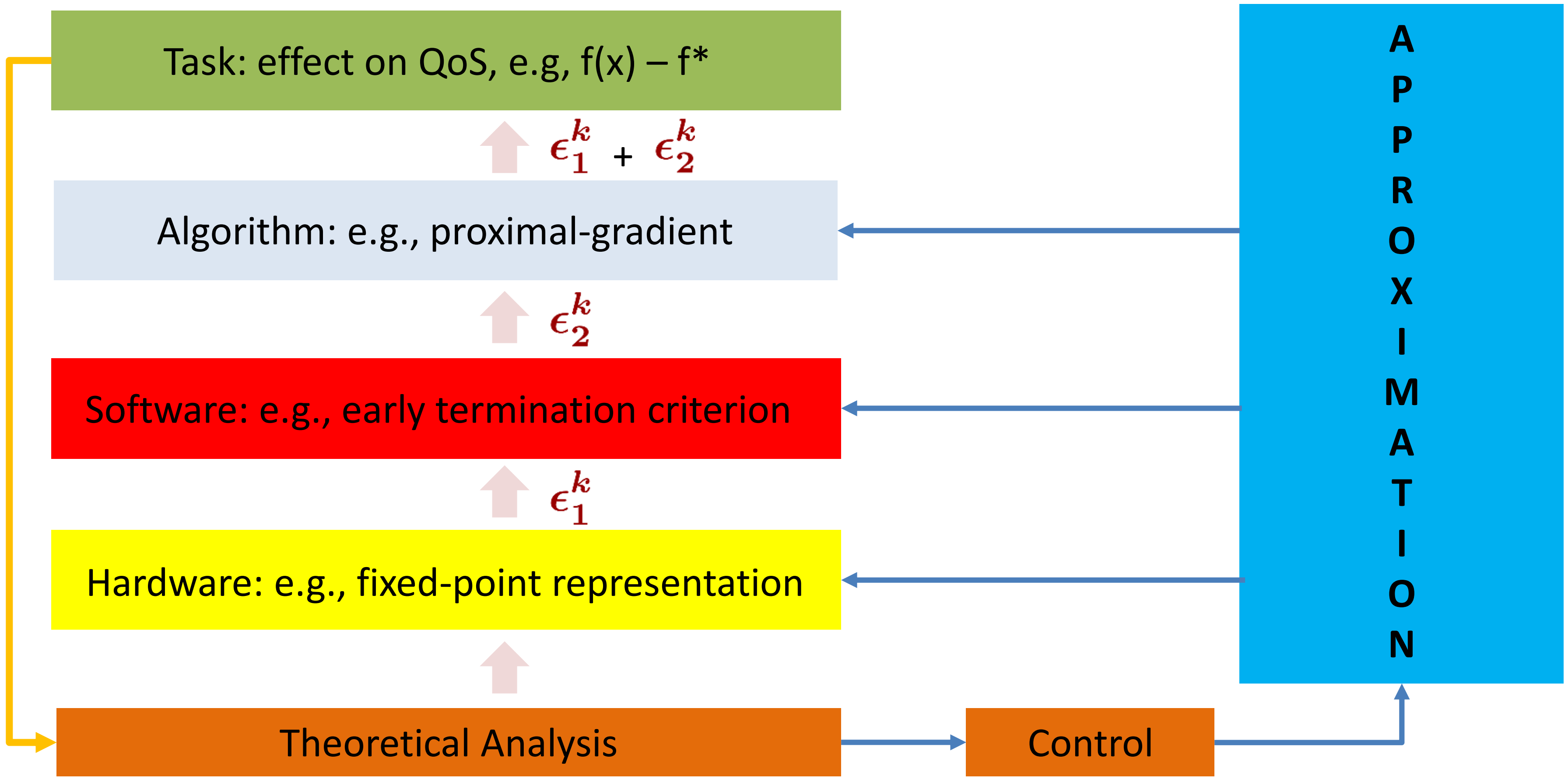
*Approximate PGD with **decreasing errors*** — *Schmidt et al. 2010*

Criterion	Bound	Constants
$f\left(\frac{1}{k}\sum_{i=1}^{k} x_i\right) - f(x^*)$	$\frac{L}{2k}\left(\|x_0 - x^*\| + 2A_k + \sqrt{2B_k}\right)^2$	$A_k = \sum_{i=1}^{k}\left(\frac{\|e_i\|}{L} + \sqrt{\frac{2\varepsilon_i}{L}}\right)$, $B_k = \sum_{i=1}^{k}\frac{\varepsilon_i}{L}$

*Approximate Accelerated PGD with **square summable (weighted) errors*** — *Schmidt et al. 2010, Aujol et al. 2015*

Criterion	Bound	Constants
$f(x_k) - f(x^*)$	$\frac{2L}{(k+1)^2}\left(\|x_0 - x^*\| + 2\widetilde{A}_k + \sqrt{2\widetilde{B}_k}\right)^2$	$\widetilde{A}_k = \sum_{i=1}^{k} i\left(\frac{\|e_i\|}{L} + \sqrt{\frac{2\varepsilon_i}{L}}\right)$, $\widetilde{B}_k = \sum_{i=1}^{k}\frac{i^2\varepsilon_i}{L}$
$t_N^2 w_N + \sum_{n=2}^{N}\rho_n w_{n-1} + \frac{1}{2\gamma}\|u_N - x^*\|^2$ $u_n = x_{n-1} + t_n(x_n - x_{n-1})$ $w_n := F(x_n) - F(x^*)$	$\frac{1}{2\gamma}\left(\|u_0 - x^*\| + 2A_{i,N} + \sqrt{2B_N}\right)^2$	$A_{1,n} = \sum_{k=1}^{n} t_k\left(\gamma\|e_k\| + \sqrt{2\gamma\varepsilon_k}\right)$ $A_{2,n} = \gamma\sum_{k=1}^{n} t_k\|e_k\|$ $B_n = \gamma\sum_{k=1}^{n} t_k^2\varepsilon_k$

Analysis of Error Propagation



Error Models

$$\underset{x\in\mathbb{R}^n}{\text{minimize}} \;\; f(x) := g(x) + h(x)$$

Error Type	Probabilistic Model	
Gradient computation (linear) $\nabla^{\epsilon_1^k} g(x^k) := \nabla g(x^k) + \epsilon_1^k$	1. Centered & CMI 2. Bounded 3. CMI of the iterates	$\mathbb{E}\Big[\epsilon_{1\Omega}^k \,\Big\vert\, \epsilon_{1\Omega}^1, \ldots, \epsilon_{1\Omega}^{k-1}\Big] = \mathbb{E}\Big[\epsilon_{1\Omega}^k\Big] = 0\,,$ $\mathbb{P}\Big(\lvert\epsilon_{1\Omega j}^k\rvert \le \delta\Big) = 1\,, \quad \text{for all } j = 1, \ldots, n,$ $\mathbb{E}\Big[{\epsilon_{1\Omega}^k}^\top x_\Omega^k \,\Big\vert\, \epsilon_{1\Omega}^1, \ldots, \epsilon_{1\Omega}^{k-1},\, x_{1\Omega}^1, \ldots, x_{1\Omega}^{k-1}\Big] = \mathbb{E}\Big[{\epsilon_{1\Omega}^k}^\top x_\Omega^k\Big] = 0\,,$
Proximal computation (nonlinear) $r^i = x^i - \bar{x}^i$ $h\big(x^{k+1}\big) + \frac{1}{2s}\Big\lVert x^{k+1} - x^k + s\nabla^{\epsilon_1^k} g(x^k)\Big\rVert_2^2 \le$ $\epsilon_2^k + h\big(\overline{x}^{k+1}\big) + \frac{1}{2s}\Big\lVert \overline{x}^{k+1} - x^k + s\nabla^{\epsilon_1^k} g(x^k)\Big\rVert_2^2$	1. Centered and CMI 2. CMI of the iterates 3. Bounded	$\mathbb{E}\Big[r_\Omega^k \,\Big\vert\, r_\Omega^1, \ldots, r_\Omega^{k-1}\Big] = \mathbb{E}\Big[r_\Omega^k\Big] = 0\,,$ $\mathbb{E}\Big[{r_\Omega^k}^\top x_\Omega^k \,\Big\vert\, r_\Omega^1, \ldots, r_\Omega^{k-1},\, x_{1\Omega}^1, \ldots, x_{1\Omega}^{k-1}\Big] = \mathbb{E}\Big[{r_\Omega^k}^\top x_\Omega^k\Big] = 0\,,$ $\mathbb{P}\Big(\lvert\epsilon_{2\Omega}^k\rvert \le \varepsilon_0\Big) = 1$

Independence ⟶ Conditional Mean Independence (CMI) ⟶ Uncorrelatedness

Probabilistic analysis is a hybrid of worst-case and average-case analyses that inherits advantages of both. It measures the expected performance of algorithms under slight random perturbations of worst-case inputs

Convergence Results

$$\boxed{\underset{x\in\mathbb{R}^n}{\text{minimize}} \;\; f(x) := g(x) + h(x)}$$

Scheme	Analysis	Bound
Proximal-Gradient Method	Deterministic	$f\Big(\frac{1}{k+1}\sum_{i=0}^{k} x^{i+1}\Big) - f(x^\star) \le \underbrace{\frac{1}{k+1}\Big[\sum_{i=0}^{k}\epsilon_2^i + \sum_{i=1}^{k}\Big(\big\|\epsilon_1^i\big\|_2+\sqrt{\frac{2\epsilon_2^i}{s}}\Big)\big\|x^\star-x^0\big\|_2}_{\text{Error term}} + \underbrace{\frac{1}{2s}\big\|x^\star-x^0\big\|_2^2}_{\text{Error-free}}\Big]$
	Probabilistic $2-4\exp(-\frac{\gamma^2}{2})$	$f\Big(\frac{1}{k}\sum_{i=1}^{k} x_\Omega^i\Big) - f(x^\star) \le \underbrace{\frac{1}{k}\sum_{i=1}^{k}\epsilon_{2\Omega}^i + \frac{\gamma}{\sqrt{k}}\Big(\sqrt{n}\lvert\delta\rvert+\sqrt{\frac{2\varepsilon_0}{s}}\Big)\big\|x^\star-x^0\big\|_2}_{\text{Error term}} + \underbrace{\frac{1}{2sk}\big\|x^\star-x^0\big\|_2^2}_{\text{Erro-free}}$

Convergence Results

$$\underset{x \in \mathbb{R}^n}{\text{minimize}} \;\; f(x) := g(x) + h(x)$$

Assumption on error for convergence	Scheme	Analysis	Error Bounds	Rate
$\epsilon_1 \propto O(1/k^{1+\lambda})$ $\epsilon_2 \propto O(1/k^{2+\lambda})$	Proximal-Gradient Method	Deterministic	$\frac{1}{k+1}\left[\sum_{i=0}^{k} \epsilon_2^i + \sum_{i=1}^{k} \left(\left\Vert \epsilon_1^i \right\Vert_2 + \sqrt{\frac{2\epsilon_2^i}{s}} \right) \left\Vert x^\star - x^0 \right\Vert_2\right.$	$O(1/k)$
$\epsilon_1 \propto O(1/k^{0.5+\lambda})$ $\epsilon_2 \propto O(1/k^{1+\lambda})$		Probabilistic $2-4\exp(-\frac{\gamma^2}{2})$	$\frac{1}{k}\sum_{i=1}^{k} \epsilon_{2_\Omega}^i + \frac{\gamma}{\sqrt{k}} \left(\sqrt{n}\lvert\delta\rvert + \sqrt{\frac{2\varepsilon_0}{s}} \right) \left\Vert x^\star - x^0 \right\Vert_2$	$O(1/k)$

Schmidt et al. 2010

Assumption on error	Scheme	Analysis	Error Bounds	Rate
$\epsilon_1 \propto O(1/k^{1+\lambda})$ $\epsilon_2 \propto O(1/k^{2+\lambda})$	Proximal-Gradient Method	Deterministic Probabilistic	$\frac{1}{2sk}\left[\left\Vert x^\star - x^0 \right\Vert_2 + 2A_k + \sqrt{2B_k}\right]^2$ $A_k = \sum_{i=1}^{k} \left(\frac{\Vert\epsilon_1^i\Vert_2}{L} + \sqrt{\frac{2\epsilon_2^i}{L}} \right), \quad B_k = \sum_{i=1}^{k} \frac{\epsilon_2^i}{L},$	$O(1/k)$

Convergence Results

$$\underset{x\in\mathbb{R}^n}{\text{minimize}} \; f(x) := g(x) + h(x)$$

Assumption on error for convergence	Scheme	Analysis	Error Bounds	Rate
$\epsilon_1 \propto O(1/k)$ $\epsilon_2 \propto O(1/k^2)$	Proximal-Gradient Method	Deterministic	$\frac{1}{k+1}\left[\sum_{i=0}^{k}\epsilon_2^i + \sum_{i=1}^{k}\left(\left\|\epsilon_1^i\right\|_2 + \sqrt{\frac{2\epsilon_2^i}{s}}\right)\left\|x^\star - x^0\right\|_2\right.$	$O(\log k/k)$
$\epsilon_1 \propto O(1/k^{0.5})$ $\epsilon_2 \propto O(1/k)$		Probabilistic $2 - 4\exp(-\frac{\gamma^2}{2})$	$\frac{1}{k}\sum_{i=1}^{k}\epsilon_{2_\Omega}^i + \frac{\gamma}{\sqrt{k}}\left(\sqrt{n}\lvert\delta\rvert + \sqrt{\frac{2\varepsilon_0}{s}}\right)\left\|x^\star - x^0\right\|_2$	$O(\log k/k)$

Schmidt et al. 2010

Assumption on error	Scheme	Analysis	Error Bounds	Rate
$\epsilon_1 \propto O(1/k)$ $\epsilon_2 \propto O(1/k^2)$	Proximal-Gradient Method	Deterministic Probabilistic	$\frac{1}{2sk}\left[\left\|x^\star - x^0\right\|_2 + 2A_k + \sqrt{2B_k}\right]^2$ $A_k = \sum_{i=1}^{k}\left(\frac{\|\epsilon_1^i\|_2}{L} + \sqrt{\frac{2\epsilon_2^i}{L}}\right), \quad B_k = \sum_{i=1}^{k}\frac{\epsilon_2^i}{L},$	$O(\log^2 k/k)$

Convergence Results

$$\boxed{\underset{x\in\mathbb{R}^n}{\text{minimize}} \;\; f(x) := g(x) + h(x)}$$

Assumption on error for convergence	Scheme	Analysis	Error Bounds	Rate
$\epsilon_1 \leq \delta$ $\epsilon_2 \leq \varepsilon_0$	Proximal-Gradient Method	Deterministic	$\frac{1}{k+1}\left[\sum_{i=0}^{k}\epsilon_2^i + \sum_{i=1}^{k}\left(\left\Vert\epsilon_1^i\right\Vert_2 + \sqrt{\frac{2\epsilon_2^i}{s}}\right)\left\Vert x^\star - x^0\right\Vert_2\right.$	$O(1)$
Stationary $\epsilon_1 \leq \delta$ $\epsilon_2 \leq \varepsilon_0$		Probabilistic $2-4\exp(-\frac{\gamma^2}{2})$	$\boxed{E\left(\epsilon_{2_\Omega}\right)} + \frac{\gamma}{\sqrt{k}}\left(\frac{\varepsilon_0}{2} + \sqrt{n}\lvert\delta\rvert\left\Vert x^\star - x^0\right\Vert_2\right)$ $E(\epsilon_{2_\Omega})$-optimality if stationary	$O(1/\sqrt{k})$

Schmidt et al. 2010

Assumption on error	Scheme	Analysis	Error Bounds	Rate
$\epsilon_1 \leq \delta$ $\epsilon_2 \leq \varepsilon_0$	Proximal-Gradient Method	Deterministic	$\frac{1}{2sk}\left[\left\Vert x^\star - x^0\right\Vert_2 + 2A_k + \sqrt{2B_k}\right]^2$ $A_k = \sum_{i=1}^{k}\left(\frac{\Vert\epsilon_1^i\Vert_2}{L} + \sqrt{\frac{2\epsilon_2^i}{L}}\right), \quad B_k = \sum_{i=1}^{k}\frac{\epsilon_2^i}{L},$	$O(k)$

Experimental Setup:

$$\underset{x\in\mathbb{R}^n}{\text{minimize}} \quad f(x) := g(x) + h(x)$$

$$x^{k+1} = \text{prox}_{\alpha_k h}\left(x^k - \alpha_k \nabla g(x^k)\right)$$

- LASSO with 600 random examples and 100 features using fixed-point representation (round-off error) and finite solver precision (CVX).

$$\underset{x\in\mathbb{R}^n}{\text{minimize}} \quad \frac{1}{2}\|Ax - y\|_2^2 + \lambda\|x\|_1 \,,$$

- Quantization according to Q-format:

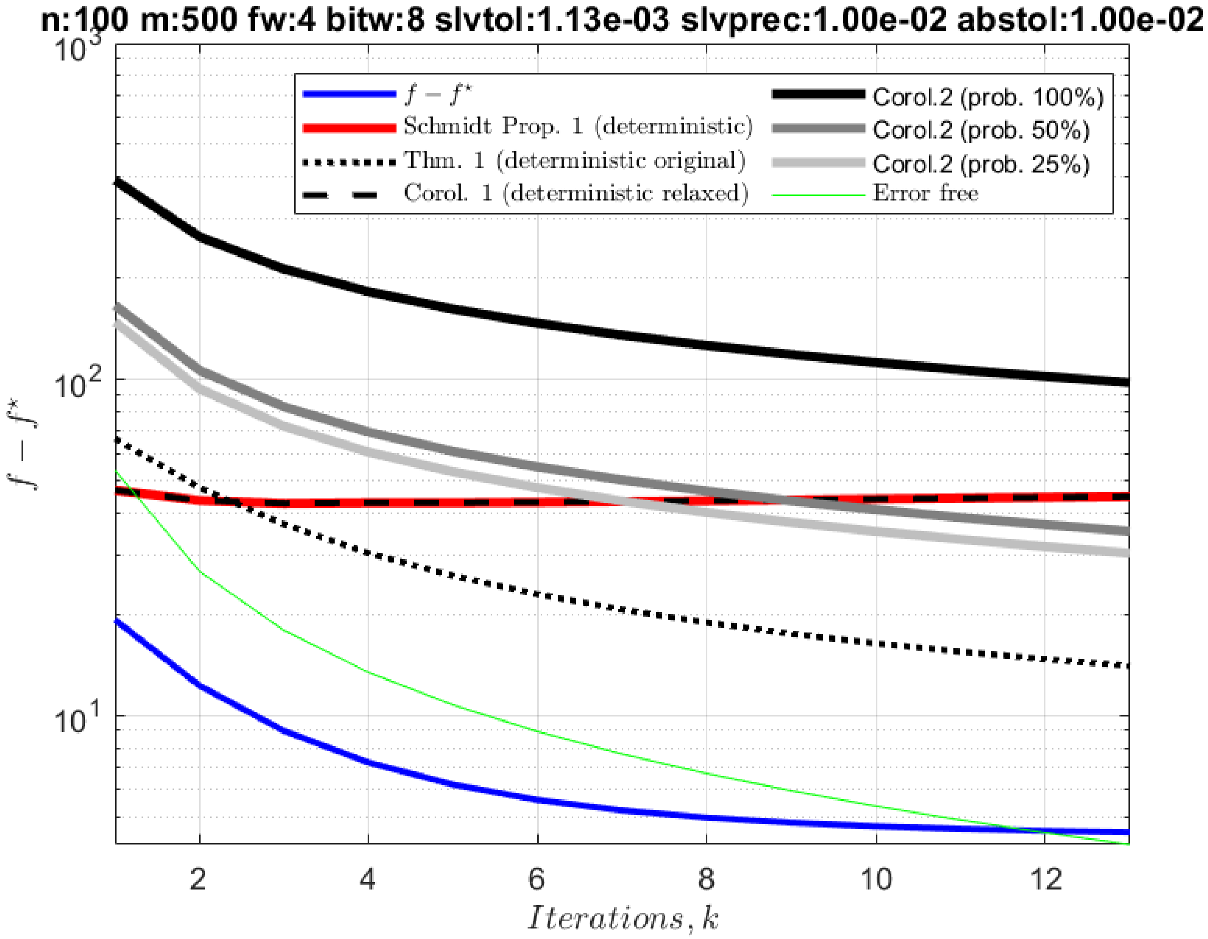
$$sI.F(x) = \sum_{i=0}^{W-2} b_i 2^{i-F} - b_{W-1} 2^{I-1}.$$

precision	bit	frac.	ABSTOL
2.22e-16	8	4	2.22e-16
			0.001
	16	6	2.22e-16
			0.01
		8	2.22e-16
0.001	8	4	0.001
			0.01
	16	8	2.22e-16
			0.001
0.01	8	4	0.01
	16	6	0.01
		8	0.01

Experimental Results:

$$\underset{x\in\mathbb{R}^n}{\text{minimize}} \;\; f(x) := g(x) + h(x)$$

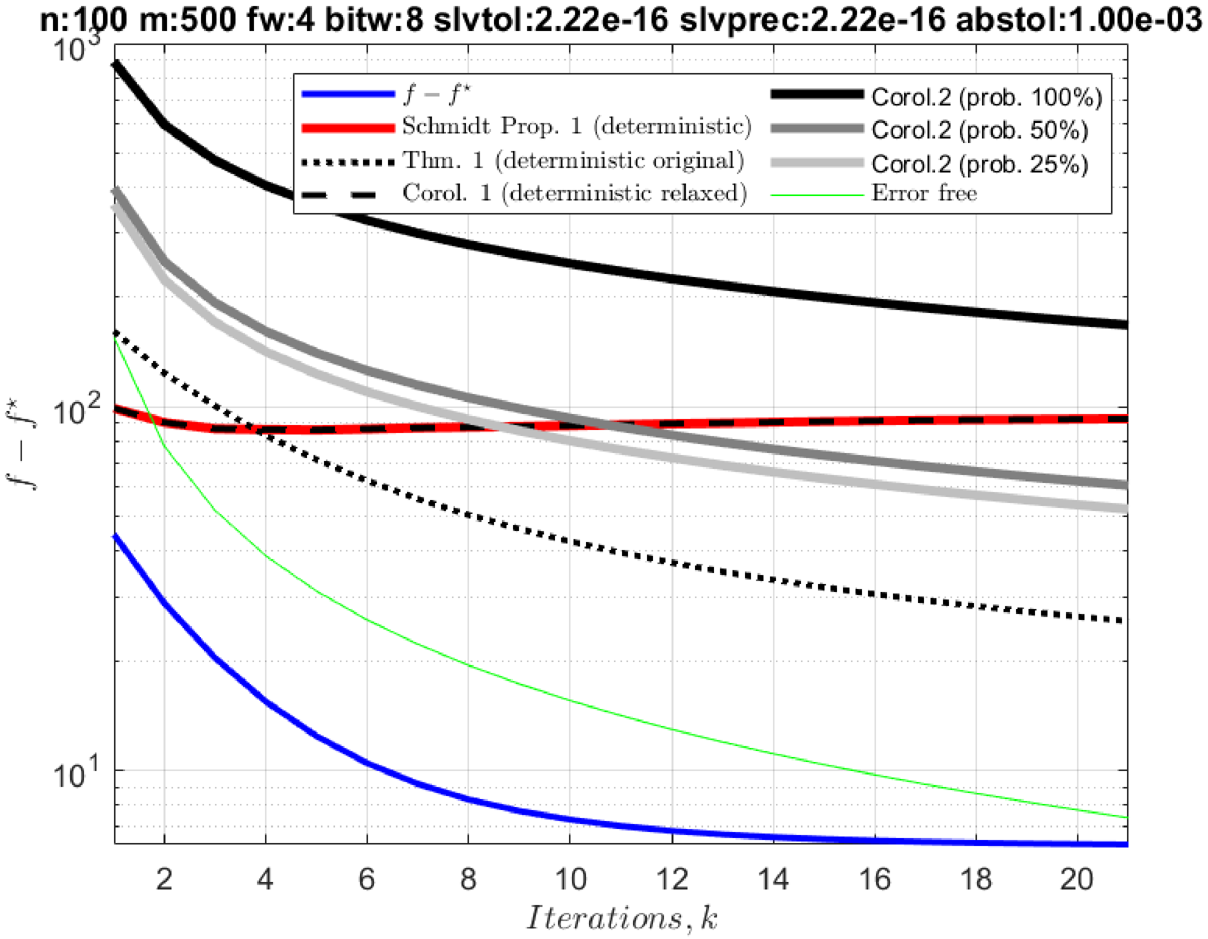
$$x^{k+1} = \text{prox}_{\alpha_k h}\Big(x^k - \alpha_k \nabla g(x^k)\Big)$$



Experimental Results:

$$\underset{x\in\mathbb{R}^n}{\text{minimize}} \quad f(x) := g(x) + h(x)$$

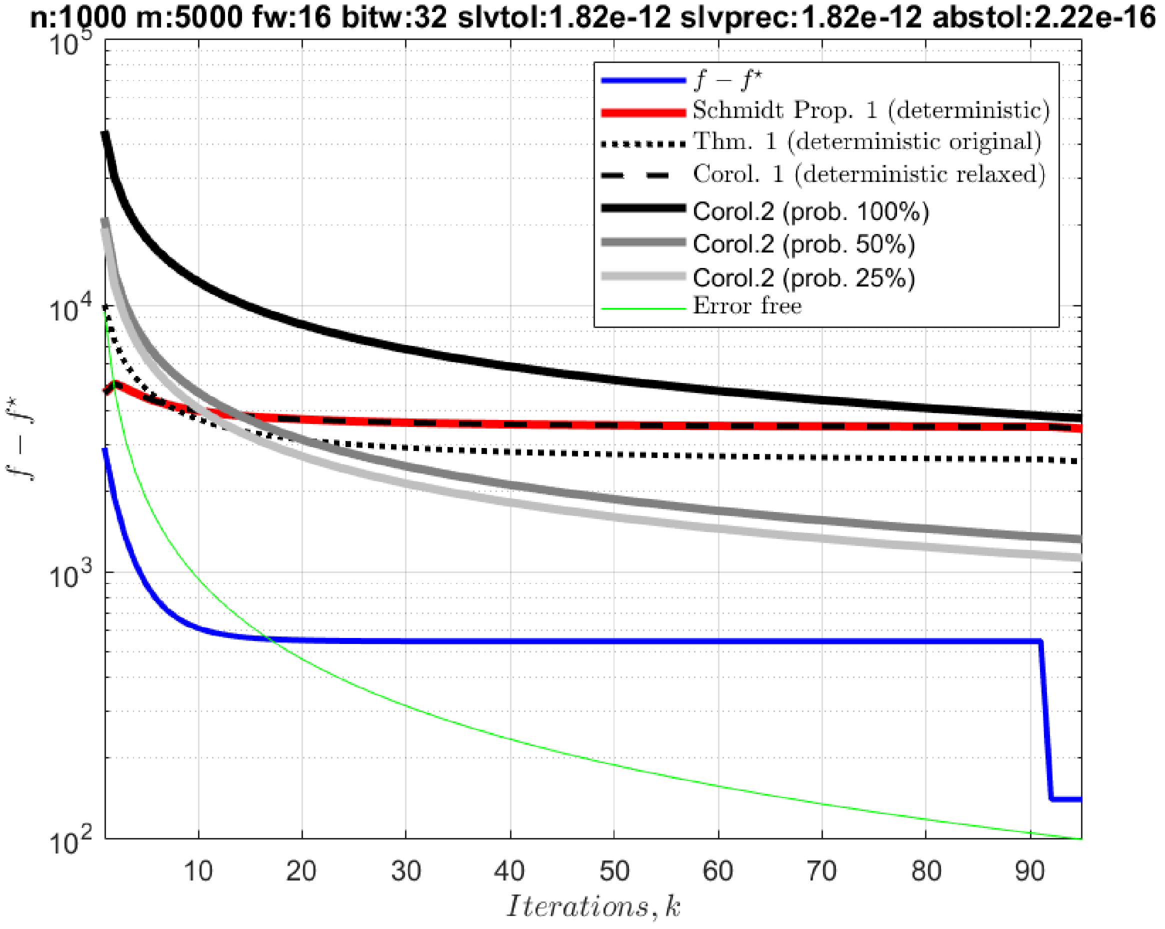
$$x^{k+1} = \text{prox}_{\alpha_k h}\Big(x^k - \alpha_k \nabla g(x^k)\Big)$$



Experimental Results:

$$\boxed{\underset{x\in\mathbb{R}^n}{\text{minimize}} \;\; f(x) := g(x) + h(x)}$$

$$x^{k+1} = \text{prox}_{\alpha_k h}\left(x^k - \alpha_k \nabla g(x^k)\right)$$



Conclusion:

- We obtained new tighter deterministic bounds and we demonstrated their validity on a practical optimization example (LASSO) solved on a reduced-precision machine combined with reduced-precision solver.

Conclusion:

- We obtained new tighter deterministic bounds and we demonstrated their validity on a practical optimization example (LASSO) solved on a reduced-precision machine combined with reduced-precision solver.
- We also derived probabilistic upper bounds.

Conclusion:

- We obtained new tighter deterministic bounds and we demonstrated their validity on a practical optimization example (LASSO) solved on a reduced-precision machine combined with reduced-precision solver.
- We also derived probabilistic upper bounds.
- Worst-case running time can be much worse than the observed running time in practice.

Conclusion:

- We obtained new tighter deterministic bounds and we demonstrated their validity on a practical optimization example (LASSO) solved on a reduced-precision machine combined with reduced-precision solver.
- We also derived probabilistic upper bounds.
- Worst-case running time can be much worse than the observed running time in practice.
- Probabilistic bounds are more practical.

Conclusion:

- We obtained new tighter deterministic bounds and we demonstrated their validity on a practical optimization example (LASSO) solved on a reduced-precision machine combined with reduced-precision solver.
- We also derived probabilistic upper bounds.
- Worst-case running time can be much worse than the observed running time in practice.
- Probabilistic bounds are more practical.
- More relaxations on the assumptions are needed in order to incorporate more general perturbations into the analysis.


thanks
...a watt!