MIAI Grenoble Alpes
Multidisciplinary Institute In Artificial Intelligence

Sparse recovery by reduced variance stochastic approximation

Anatoli Juditsky, *Université Grenoble-Alpes*

based on the joint work with Andrei Kulunchakov and Hlib Tsyntseus

OWB, Grenoble-Sochi, July 16, 2021

Supported by MIAI@Grenoble Alpes (ANR-19-P3IA-0003)

MOTIVATION

Consider the problem of recovery of an unknown signal $x_* \in R^n$ from linear noisy observations (linear regression):

$$\eta_i = \phi_i^T x_* + \sigma\xi, \quad i = 1,...,N,$$

where $\phi_i \in R^n$ are "random regressors" and ξ_i are zero-mean noises with $E\{\xi_i^2\} \leq 1$; we suppose that ϕ_i and ξ_i are i.i.d..

- In our setting, n is "large" and

$$N \ll n,$$

but x_* is s-sparse, namely, has $\leq s$ nonvanishing entries.

- Let us consider Stochastic Optimization problem

$$\min_{x \in X} \left\{ g(x) = E\{\tfrac{1}{2}(\eta_1 - \phi_1^T x)^2\} =: E\{G(x, \omega_i = [\phi_i, \eta_i])\} \right\}. \qquad \text{(SR)}$$

We assume that $E\{\phi_i \phi_i^T\} = \Sigma \succ 0$,

$$g(x) = E\{G(x, \omega_i)\} = \tfrac{1}{2}(x - x_*)^T \Sigma (x - x_*) + \sigma^2,$$

and x_* is the unique minimizer of (SR).

$= \ldots = n - (1 - \gamma^{-2}) \sum_{i=0}^{k} \| H_i g_i \|^2 / \| g_i \|^2 > 0.$

Положим

$v_i = \langle g_i , x_i - x^* \rangle / \| g_i \|$

Тогда из предыдущего неравенства и неравенства (3.5.5) получаем

$\sum_{i=0}^{k} v_i^4 \leq \sum_{i=0}^{k} \| g_i \|^{-4} \langle H_i g_i , g_i \rangle^2 \times$

$\times \langle G_i (x_i - x^*) , x_i - x^* \rangle^2 \leq$

$\leq r^4 \sum_{i=0}^{k} \| g_i \|^{-4} \langle H_i g_i , g_i \rangle^2 \leq$

$\leq r^4 \sum_{i=0}^{k} \| g_i \|^{-2} \| H_i g_i \|^2 \leq$

$\leq n \gamma^2 (\gamma^2 - 1)^{-1} r^4 .$

Отсюда, воспользовавшись следствием 1.5.1, получаем, что $f(x_k) \to f^*$ при $k \to \infty$

Далее,

$$\det G_{k+1} = \det \left[G_k \left(I + (\gamma^2 - 1) \frac{H_k g_k g_k^T}{\langle H_k g_k , g_k \rangle} \right) \right] =$$

$= \gamma^2 \det G_k = \ldots = \gamma^{2(k+1)} .$

Положим $b_k = \| g_k \|^{-2} \langle H_k g_k , g_k \rangle$. Тогда

$\gamma^{2(k+1)} = n [\det G_{k+1}]^{1/n} \leq \mathrm{Trace}\, G_{k+1} =$

$= n + (\gamma^2 - 1) \sum_{i=0}^{k} b_i^{-1} \leq n + (b_k^*)^{-1} (\gamma^2 - 1)(k + 1)$

Таким образом,

$b_k^* \leq n^{-1} (\gamma^2 - 1)(k + 1) [\gamma^{2(k+1)/n} - 1]^{-1}$

для всех $k > 0$. Следовательно,

$(v_k^*)^2 \leq b_k^* r^2 \leq$

$\leq n^{-1} (\gamma^2 - 1)(k + 1) [\gamma^{2(k+1)/n} - 1]^{-1} r^2$

Для завершения доказательства осталось воспользоваться следствиями 3.5.1 и 1.5.1. □

ГЛАВА 4. МЕТОДЫ РЕШЕНИЯ ЭКСТРЕМАЛЬНЫХ ЗАДАЧ С ГЛАДКИМИ КОМПОНЕНТАМИ

4.1. Оптимальные методы безусловной минимизации функций с липшицевым градиентом

В этом параграфе рассматриваются итеративные методы решения следующей экстремальной задачи:

$$\min \{ f(x) \mid x \in R^n \} , \qquad (4.1.1)$$

где $f(x)$ - выпуклая функция из класса $\mathcal{F}(L , m)$.

Прежде всего остановимся на способе получения оценок скорости сходимости, который будет использоваться в настоящем параграфе.

Пусть $\{ x_k \}_{k=0}^{\infty}$ - последовательность точек, вырабатываемая методом $\mathcal{M}$, $\Psi_k (x)$, $x \in R^n$, - последовательность функций таких, что для любого x из R^n

$$\lim_{k \to \infty} \Psi_k (x) = 0 .$$

Предположим, что при любом $k > 0$ и $x \in R^n$ справедливо неравенство

$$f(x_k) \leq f(x) + \Psi_k(x) . \qquad (4.1.2)$$

Тогда очевидно,

$$f(x_k) - f^* \leq \Psi_k(x^*) \to 0 \quad \text{при} \quad k \to \infty \qquad (4.1.3)$$

Таким образом, если метод $\mathcal{M}$ обеспечивает выполнение неравенства (4.1.2), то в силу (4.1.3) оценка стремления к нулю числовой последовательности $\{ \Psi_k (x^*) \}$ дает глобальную оценку скорости сходимости метода $\mathcal{M}$.

Способ (4.1.2),(4.1.3) при соответствующем выборе функций Ψ_k позволяет получать оценки скорости сходимости некоторых известных методов, например метода градиентного спус-

Motivation

Consider the problem of recovery of an unknown signal $x_* \in \mathbb{R}^n$ from linear noisy observations (linear regression):

$$\eta_i = \phi_i^T x_* + \sigma\xi, \quad i = 1, ..., N,$$

where $\phi_i \in \mathbb{R}^n$ are "random regressors" and ξ_i are zero-mean noises with $\mathbf{E}\{\xi_i^2\} \leq 1$; we suppose that ϕ_i and ξ_i are i.i.d..

- In our setting, n is "large" and

$$N \ll n,$$

but x_* is s-sparse, namely, has $\leq s$ nonvanishing entries.

- Let us consider Stochastic Optimization problem

$$\min_{x \in X} \left\{ g(x) = \mathbf{E}\{\tfrac{1}{2}(\eta_1 - \phi_1^T x)^2\} =: \mathbf{E}\{G(x, \omega_i = [\phi_i, \eta_i])\} \right\}. \tag{SR}$$

We assume that $\mathbf{E}\{\phi_i \phi_i^T\} = \Sigma \succ 0$,

$$g(x) = \mathbf{E}\{G(x, \omega_i)\} = \tfrac{1}{2}(x - x_*)^T \Sigma (x - x_*) + \sigma^2,$$

and x_* is the unique minimizer of (SR).

There are several approaches to solving (SR).

• Note that observations η_i and ϕ_i provide us with unbiased estimates $G(x,\omega_i)$ of the problem objective $g(x)$, so one can build a Sample Average Approximation (SAA)

$$\widehat{g}(x) = \frac{1}{N}\sum_{i=1}^{N} G(x,\omega_i) = \tfrac{1}{2N}\|\eta - \Phi^T x\|_2^2, \qquad \Phi = [\phi_1,...,\phi_N],$$

of $g(x)$ and then solve the the problem by a deterministic optimization routine.

• Iterative thresholding algorithms *Blumensath, Davies '09, Foygel Barber et al. '18, '19, Jain et al. '14, ...*

• ℓ_1-minimization—replacing sparsity constrained minimization with ℓ_1-penalization, e.g., solving instead

$$\min_x \widehat{g}(x) + \kappa\|x\|_1, \qquad \kappa > 0, \qquad \text{(Lasso)}$$

Bickel et al. '09, Candes et al. '06, '07,..., Dalalyan, Thompson '19, Fazel '08, ...

• Let $\|a\|_{s,1}$ be the sum of s largest amplitudes of a.

Assume that $\Phi = [\phi_1, ..., \phi_N]$ satisfies

$$\|z\|_{s,1} \leq \lambda\sqrt{s}\|\Phi^T z\|_2 + \chi\|z\|_1. \qquad (Q(\chi,\lambda))$$

When penalty κ is chosen properly and condition $Q(\chi,\lambda)$ holds with $\chi < \frac{1}{2}$, for any s-sparse $x_* \in \mathbb{R}^n$, solution $\widehat{x}_N$ to (Lasso) satisfies "with high probability"

$$\|\widehat{x}_N - x_*\|_1 \lesssim \tau\frac{\lambda^2 s\sigma}{\sqrt{N}}$$

with τ containing "logarithmic factors" in N and n.

• For certain distributions of ϕ's, matrix Φ satisfies $Q(\chi,\lambda)$ with $\chi < 1/2$ for $s \asymp m/\ln(n/N)$ with "high probability."

Utilizing Stochastic Approximation (SA)

Agarval et al. '12, Gaillard, Wintenberger '17, Nguyen et al. '17, Shalev-Shwartz et al. '11, Srebro et al. '10, ...

Note that

$$\nabla G(x, \omega_i) = \phi_i \phi_i^T (x - x_*) - \sigma \xi_i \phi_i$$

is an unbiased estimate of $\nabla g(x) = \Sigma(x - x_*)$, with

$$\zeta(x, \omega_i) = \nabla G(x, \omega_i) - \nabla g(x) = (\phi_i \phi_i^T - \Sigma)(x - x_*) - \sigma \xi_i \phi_i.$$

- Consider a toy situation in which
  - regressors ϕ_i are a.s. bounded, $\|\phi_i\|_\infty \leq r < \infty$ with identity covariance matrix $\Sigma = \mathbf{E}\{\phi_1 \phi_1^T\} = I$
  - noise variance σ^2 is "small"

    ("in the limit" we are looking for a sparse solution of the system $\Phi^T x = \eta$)
  - we know that $\|x_*\|_1 \leq R$, i.e., $x_* \in X = \{x \in \mathbb{R}^n : \|x\|_1 \leq R\}$.

- When using "standard" (Euclidean) Stochastic Approximation,

$$x_t = \pi_X\big[x_{t-1} - \gamma_t \nabla G(x_{t-1}, \omega_t)\big], \quad x_0 = 0$$

(here π_X is the Euclidean projection on X and $\gamma_t > 0$ are "appropriate" stepsizes), after $N \gg 1$ steps one has

$$\mathbf{E}\{g(x_N)\} - g_* \asymp \frac{\mathbf{E}\{\|\zeta(x_*, \omega_1)\|_2^2\}}{N}$$

where x_N is the approximate solution by SA, with

$$\mathbf{E}\{\|\zeta(x_*, \omega_1)\|_2^2\} = \sigma^2 \mathbf{E}\{\|\phi_1\|_2^2\} \asymp \sigma^2 n$$

leading to the error estimate

$$\mathbf{E}\{g(x_N)\} - g_* \asymp \sigma^2 \frac{n}{N}$$

depending on the problem dimension n.

• (Non-Euclidean) Stochastic Mirror Descent algorithm allows to "remove" the "n-factor".

Using "ℓ_1-Mirror Descent" with constant stepsize parameter $\beta \geq 1$ we get (up to logarithmic in n factors)

$$\sum_{t=1}^{N}[\mathbf{E}\{g(x_t)\} - g_*] \lesssim R^2\beta + \beta^{-1}\sum_{t=1}^{N}\mathbf{E}\{\|\zeta(x_{t-1},\omega_t)\|_\infty^2\}. \qquad (\mathrm{MD}_1)$$

Now

$$\begin{array}{rcl}\varsigma(x) := \mathbf{E}\{\|\zeta(x,\omega)\|_\infty^2\}^{1/2} & = & \mathbf{E}\{\|(\phi\phi^T - I)(x - x_*)\|_\infty^2\}^{1/2} + \sigma^2\mathbf{E}\{\|\xi\phi\|_\infty^2\}^{1/2} \\ & \leq & \mathbf{E}\{\|\phi\|_\infty^2[\phi^T(x - x_*)]^2\}^{1/2} + \sigma^2\mathbf{E}\{\|\phi\|_\infty^2\}^{1/2} \\ & \leq & r^2\|x - x_*\|_1 + \sigma r \leq r^2R + \sigma r\end{array}$$

does not depend on dimension; choosing $\beta = (r^2 + \sigma r/R)\sqrt{N}$, we obtain by convexity of g:

$$N\left[\mathbf{E}\left\{g\Big(\underbrace{\frac{1}{N}\sum_{t=1}^{N}x_t}_{\overline{x}_N}\Big)\right\} - g_*\right] \leq \sum_{t=1}^{N}[\mathbf{E}\{g(x_t)\} - g_*] \lesssim [rR^2 + \sigma R]N^{1/2},$$

so that

$$\mathbf{E}\left\{g\left(\overline{x}_N\right)\right\} - g_* \lesssim \frac{r^2R^2 + \sigma rR}{\sqrt{N}}.$$

Question: *how this bound can be improved?*

• (Non-Euclidean) Stochastic Mirror Descent algorithm allows to "remove" the "n-factor".

Using "ℓ_1-Mirror Descent" with constant stepsize parameter $\beta \geq 1$ we get (up to logarithmic in n factors)

$$\sum_{t=1}^{N}[\mathbf{E}\{g(x_t)\} - g_*] \lesssim R^2\beta + \beta^{-1}\sum_{t=1}^{N}\mathbf{E}\{\|\zeta(x_{t-1},\omega_t)\|_\infty^2\}. \qquad (\mathrm{MD}_1)$$

Now

$$\begin{array}{rcl}\varsigma(x) := \mathbf{E}\{\|\zeta(x,\omega)\|_\infty^2\}^{1/2} & = & \mathbf{E}\{\|(\phi\phi^T - I)(x - x_*)\|_\infty^2\}^{1/2} + \sigma^2\mathbf{E}\{\|\xi\phi\|_\infty^2\}^{1/2} \\ & \leq & \mathbf{E}\{\|\phi\|_\infty^2[\phi^T(x - x_*)]^2\}^{1/2} + \sigma^2\mathbf{E}\{\|\phi\|_\infty^2\}^{1/2} \\ & \leq & r^2\|x - x_*\|_1 + \sigma r \leq r^2R + \sigma r\end{array}$$

does not depend on dimension; choosing $\beta = (r^2 + \sigma r/R)\sqrt{N}$, we obtain by convexity of g:

$$N\left[\mathbf{E}\left\{g\Big(\underbrace{\frac{1}{N}\sum_{t=1}^{N}x_t}_{\overline{x}_N}\Big)\right\} - g_*\right] \leq \sum_{t=1}^{N}[\mathbf{E}\{g(x_t)\} - g_*] \lesssim [rR^2 + \sigma R]N^{1/2},$$

so that

$$\mathbf{E}\left\{g\left(\overline{x}_N\right)\right\} - g_* \lesssim \frac{R^2}{N} + \frac{r^2R^2 + \sigma rR}{\sqrt{N}}.$$

Question: *how this bound can be improved?*

- Using "strong convexity" of g,

$$g(x) - g_* = \tfrac{1}{2}\|x - x_*\|_2^2.$$

We want to use strong convexity w.r.t. $\|\cdot\|_2$ in the non-Euclidean algorithm tuned for $\|\cdot\|_1$

$\Rightarrow$ Use sparsity: if x is s-sparse, one has

$$\|x - x_*\|_1 \leq \sqrt{2s}\|x - x_*\|_2,$$

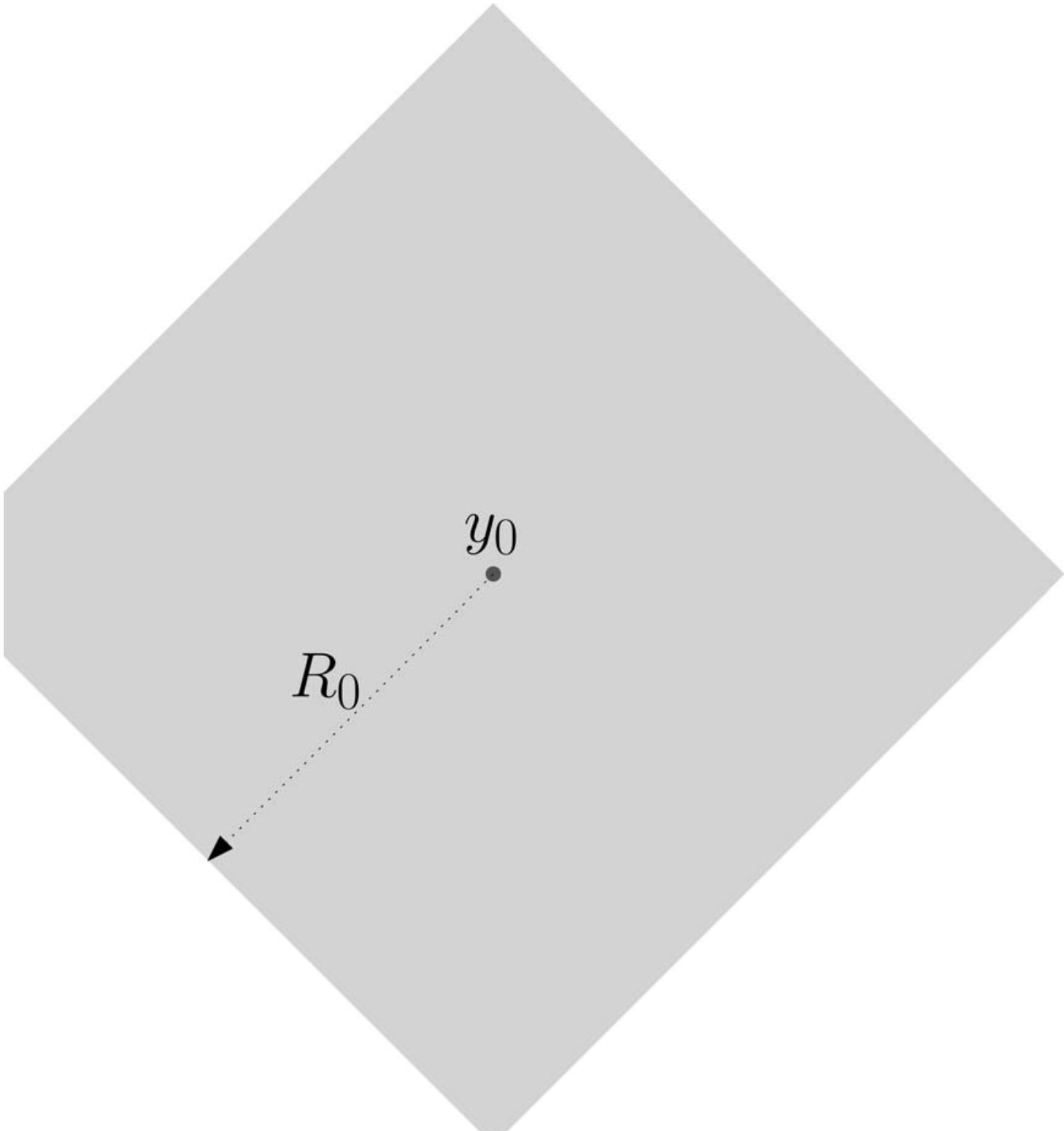
thus

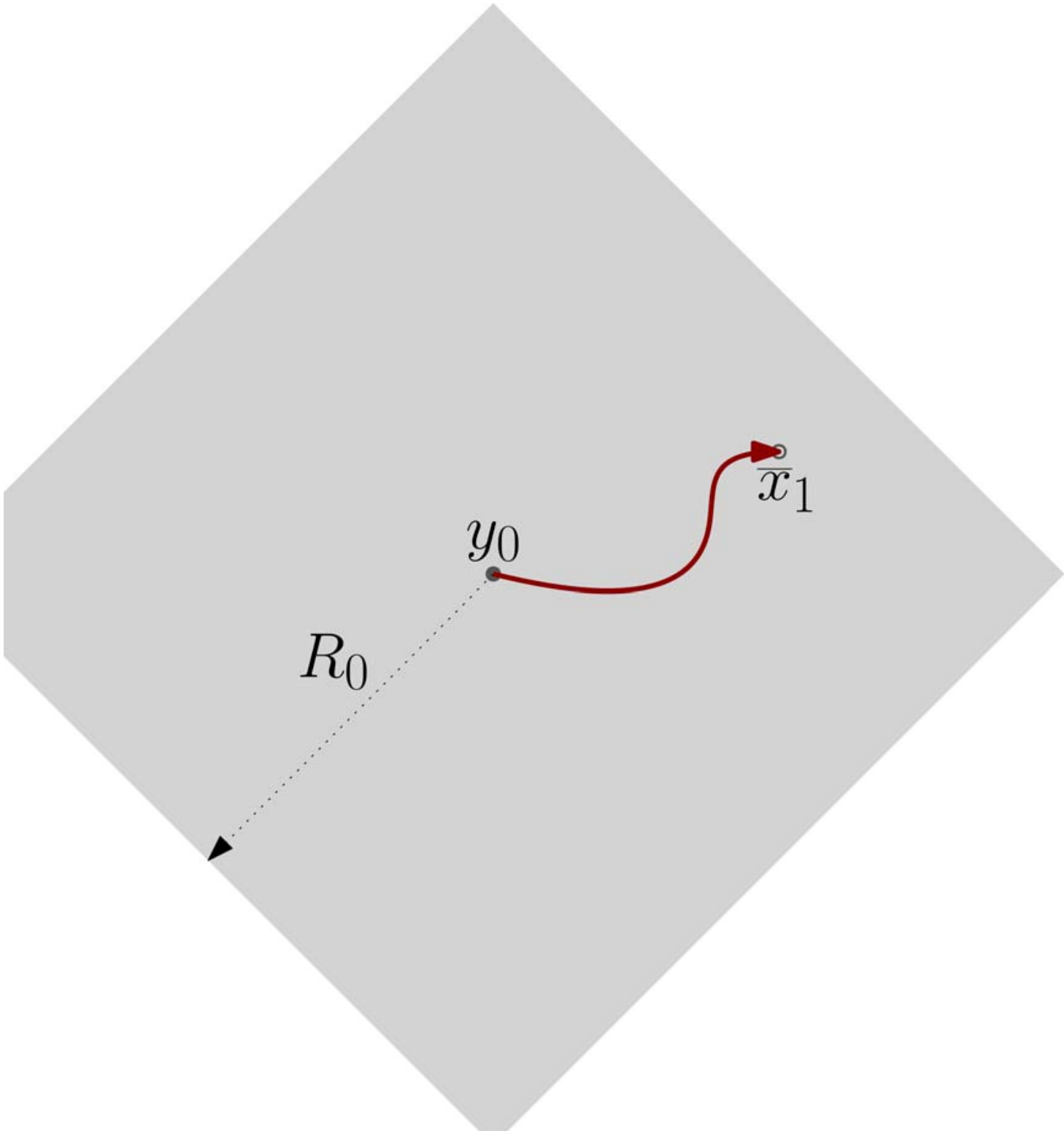
$$g(x) - g_* = \tfrac{1}{2}\|x - x_*\|_2^2 \geq \frac{1}{4s}\|x - x_*\|_1^2.$$

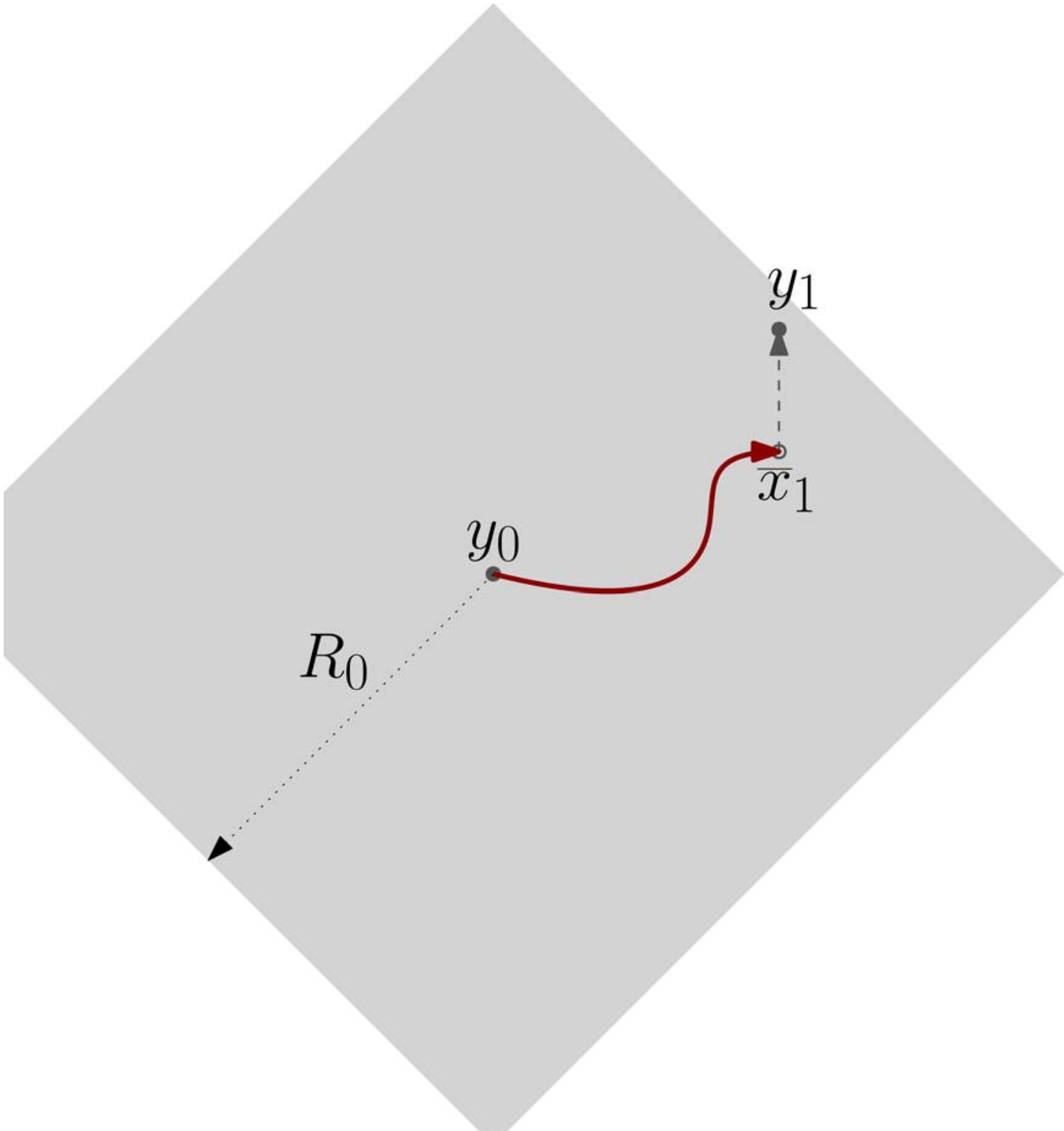
$\Rightarrow$ Organize the algorithm "in stages":

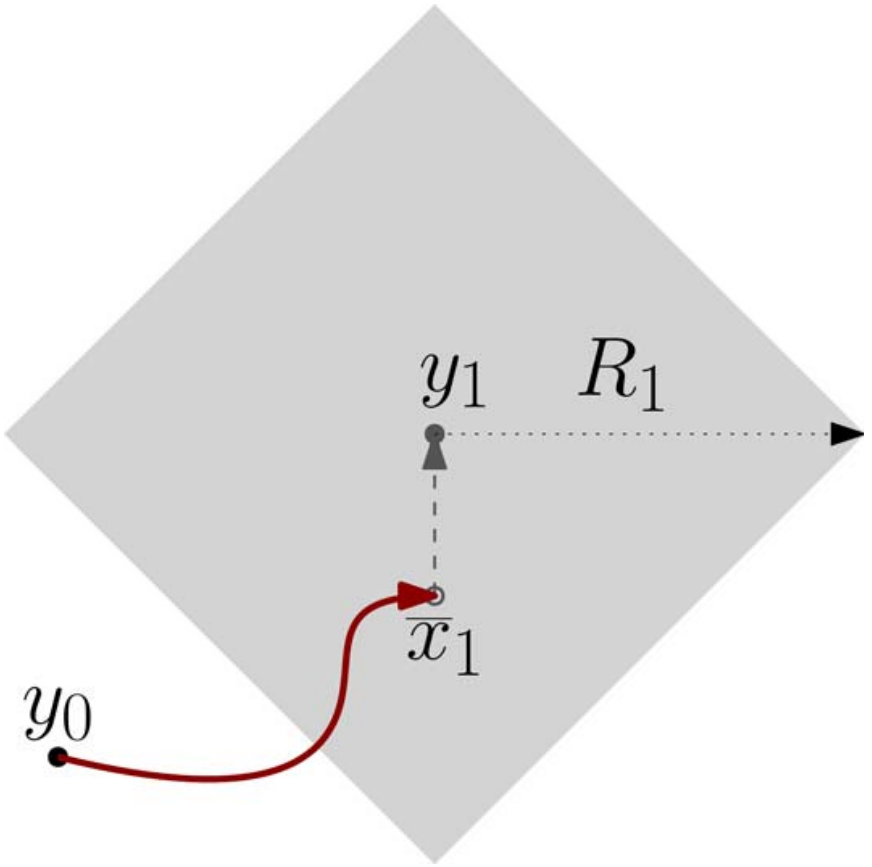
— at the k-th stage of the method, run N_k iterations of the Stochastic Mirror Descent recursion

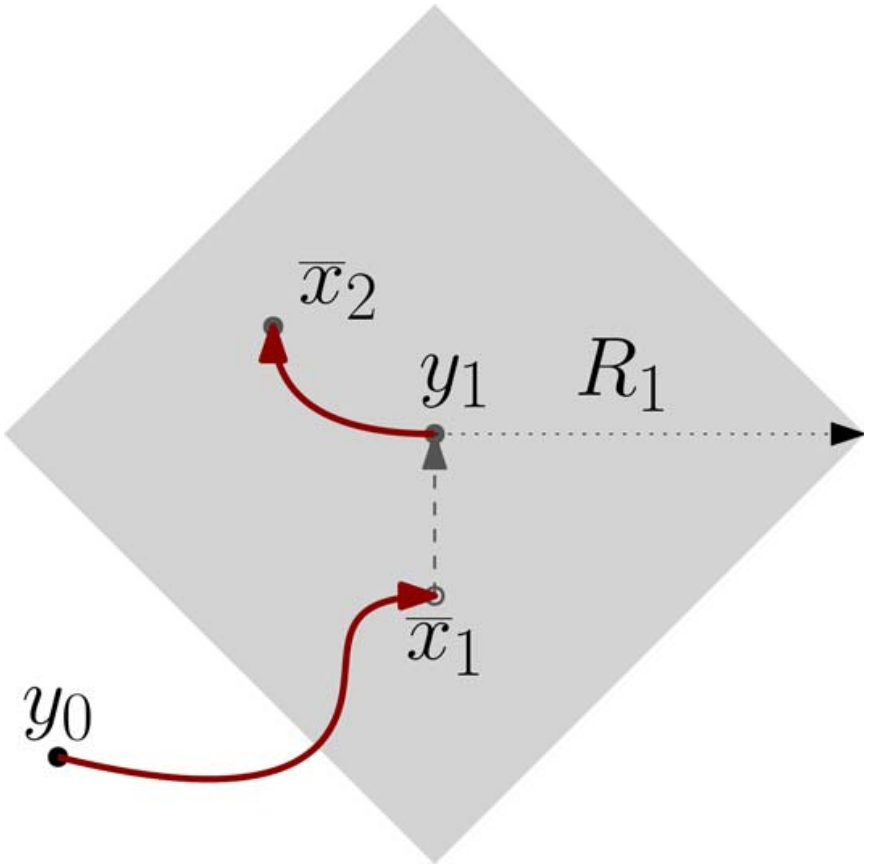
— then, "sparsify" the obtained approximate solution by zeroing out all but s entries of largest amplitudes and use strong convexity to update the bound for the error of solution.

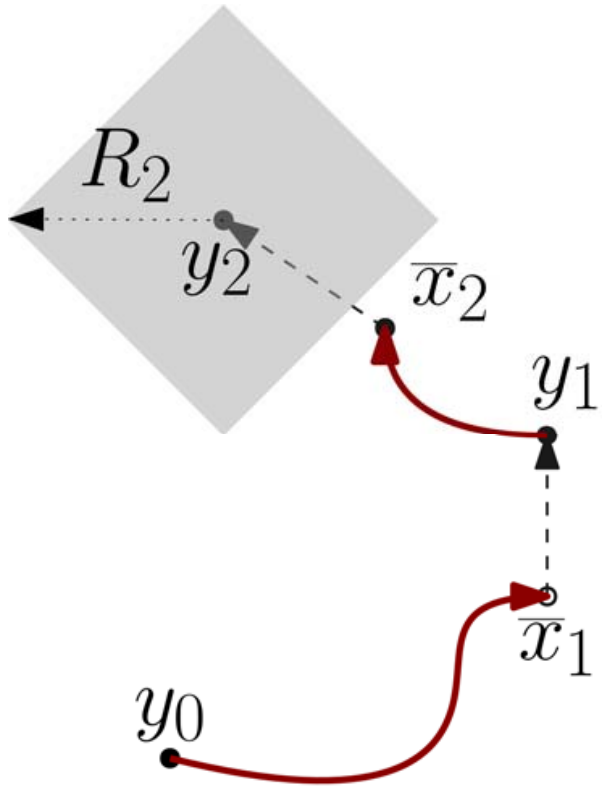

y_0
R_0

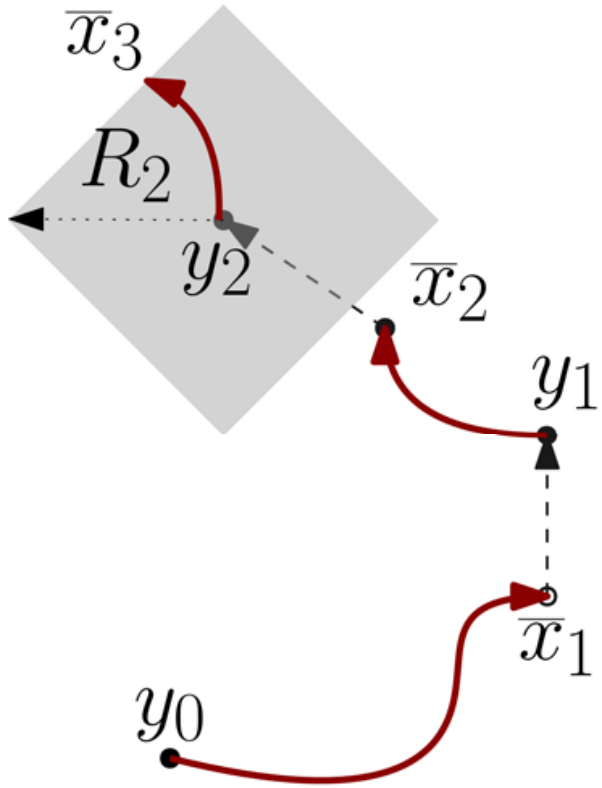

$\overline{x}_1$
y_0
R_0


y_1
$\overline{x}_1$
y_0
R_0


y_1
R_1
$\overline{x}_1$
y_0


$\overline{x}_2$
y_1
R_1
$\overline{x}_1$
y_0


R_2
y_2
$\overline{x}_2$
y_1
$\overline{x}_1$
y_0


$\overline{x}_3$
R_2
y_2
$\overline{x}_2$
y_1
$\overline{x}_1$
y_0

- Refine error bounds of Mirror Descent.

Note that the error $\zeta(x,\omega_1)$ of the stochastic gradient can be decomposed into

$$\zeta(x,\omega_1) = \underbrace{[\phi_1\phi_1^T - I](x-x_*)}_{=:\zeta_1(x,\omega_1)} + \underbrace{\sigma\xi_1\phi_1}_{=:\zeta_2(\omega_1)}.$$

"Variance" $\varsigma_1^2(x)$ of the first component is proportional to $g(x) - g(x_*)$:

$$\varsigma_1^2(x) = \mathbf{E}\{\|\zeta_1(x,\omega)\|_\infty^2\} \le 2(r^2+1)\|x-x_*\|_2^2 \lesssim r^2[g(x)-g_*],$$

while "variance" ς_2^2 of the second,

$$\varsigma_2^2 = \mathbf{E}\{\|\zeta_2(\omega)\|_\infty^2\} \le \sigma^2 r^2$$

does not depend on x (and is small when σ^2 is small).
As a result, for the "total" variance $\varsigma^2(x)$ we get

$$\varsigma^2(x) = \mathbf{E}\{\|\zeta(x,\omega)\|_\infty^2\} \lesssim r^2[g(x)-g_*] + \sigma^2 r^2.$$

When submitting into (MD_1) we get

$$\sum_{t=1}^{N}[\mathbf{E}\{g(x_t)\} - g_*] \lesssim R^2\beta + \beta^{-1}\sum_{t=1}^{N}\underbrace{(r^2[g(x_{t-1}) - g_*] + \sigma^2 r^2)}_{\gtrsim \mathbf{E}\{\|\zeta(x_{t-1},\omega_t)\|_\infty^2},$$

and for $\beta > r^2$,

$$\sum_{t=1}^{N}[\mathbf{E}\{g(x_t)\} - g_*] \lesssim R^2\beta + \frac{r^2[g(x_0) - g_*]}{\beta} + \frac{N\sigma^2 r^2}{\beta},$$

what results in

$$\mathbf{E}\left\{g\left(\overline{x}_N\right)\right\} - g_* \leq \frac{\beta R^2}{N} + \frac{r^4 R^2}{\beta N} + \frac{\sigma^2 r^2}{\beta}.$$

$\Rightarrow$ Faster convergence in the situation of "small additive noise" (small σ).

Problem setting

• Let E be a Euclidean space. Consider a Stochastic Optimization problem

$$\min_{x\in X}\left[\mathbf{E}\{G(x,\omega)\}\right] \tag{SO}$$

where

— $X \subset E$ is a convex set with nonempty interior

— ω is a random variable on a probability space Ω with distribution P

— $G:\ X\times\Omega\to\mathbb{R}$

• Let $\|\cdot\|$ be a norm on E, and let $\|\cdot\|_*$ be the conjugate norm, i.e.,

$$\|s\|_* = \max_x\{\langle s,x\rangle:\ \|x\|\le 1\},\qquad s\in E.$$

We suppose that

— the expected objective $g(x) = \mathbf{E}\{G(x,\omega)\}$ is finite for all $x\in X$, convex and differentiable on X with Lipschitz-continuous on X gradient $\nabla g(\cdot)$:

$$\|\nabla g(x') - \nabla g(x)\|_* \le \mathcal{L}\|x-x'\|,\qquad \forall\, x,x'\in X. \tag{Lip}$$

— the problem is solvable and $g(\cdot)$ satisfies the quadratic growth condition on X w.r.t. $\|\cdot\|_2$:

$$\forall x\in X:\ \ g(x)-g(x_*)\ge \tfrac{1}{2}\underline{\kappa}\|x-x_*\|_2^2.$$

• We assume that we have access to a stochastic (gray box) oracle—a device which can generate $\omega \sim P$ and compute $\forall x \in X$ a random unbiased estimation

$$\nabla G(x,\omega)[:= \nabla_x G(x,\omega)]$$

of $\nabla g(x)$.

Assumption [S1]. *$G(\cdot,\omega)$ is smooth on X, i.e., it is continuously differentiable on X for almost all $\omega \in \Omega$, and*

$$\|\nabla G(x,\omega) - \nabla G(x',\omega)\|_* \le \mathcal{L}(\omega)\|x - x'\|$$

with $\mathbf{E}\{\mathcal{L}(\omega)\} \le \nu < \infty$. We assume that

$$\mathbf{E}\{\nabla G(x,\omega)\} = \nabla g(x), \quad \mathbf{E}\{\| \underbrace{\nabla G(x,\omega) - \nabla g(x)}_{=:\zeta(x,\omega)} \|_*^2\} \le \varsigma^2(x), \quad \forall\, x \in X,$$

and, furthermore, there are $1 \le \varkappa, \varkappa' < \infty$ such that the bound holds:

$$\varsigma^2(x) \le \varkappa\nu[g(x) - g(x_*) - \langle \nabla g(x_*), x - x_*\rangle] + \varkappa' \underbrace{\mathbf{E}\{\|\zeta(x_*,\omega)\|_*^2\}}_{=:\varsigma_*^2}. \tag{S1}$$

• In the case of toy linear regression example above, stochastic gradient

$$\nabla G(x,\omega) = \phi\phi^T(x - x_*) + \sigma\xi\phi$$

satisfies Assumption S1 with $\varsigma_*^2 = \sigma^2 r^2$, $\nu = r^2 + 1$, $\kappa = 8$ and $\kappa' = 2$.

• More generally, when stochastic gradient $\nabla G(\cdot,\omega)$ is Lipschitz continuous with a.s. bounded Lipschitz constant $\mathcal{L}(\omega) \le \nu$,

$$\begin{aligned}\varsigma^2(x) &= \mathbf{E}\{\|\nabla G(x,\omega) - \nabla g(x)\|_*^2\} \le \Big(\mathbf{E}\{\|\nabla G(x,\omega) - \nabla G(x_*,\omega)\|_*^2\}^{1/2} \\ &\quad + \|\nabla g(x) - \nabla g(x_*)\|_* + \underbrace{\mathbf{E}\{\|\nabla G(x_*,\omega) - \nabla g(x_*)\|_*^2\}}_{=\varsigma_*^2}{}^{1/2}\Big)^2.\end{aligned}$$

On the other hand, by Lipschitz continuity of $\nabla G(\cdot,\omega)$ we have

$$G(x,\omega) - G(x_*,\omega) \ge \langle \nabla G(x_*,\omega), x - x_*\rangle + (2\nu)^{-1}\|\nabla G(x,\omega) - \nabla G(x_*,\omega)\|_*^2,$$

implying that

$$\begin{aligned}\varsigma^2(x) &\le \Big([2\nu\mathbf{E}\{G(x,\omega) - G(x_*,\omega) - \langle \nabla G(x_*,\omega), x - x_*\rangle\}]^{1/2} \\ &\quad + [2\nu(g(x) - g(x_*) - \langle \nabla g(x_*), x - x_*\rangle)]^{1/2} + \varsigma_*\Big)^2 \\ &\lesssim \nu[g(x) - g(x_*) - \langle \nabla g(x_*), x - x_*\rangle] + \varsigma_*^2.\end{aligned}$$

Assumption [S2] *The optimal solution x_* to problem (SO) is s-sparse.*

Furthermore, given $x \in X$ one can efficiently compute a "sparse approximation" of x—an optimal solution $x_s = \mathrm{sparse}(x)$ to the optimization problem

$$\min \|x - z\|_2 \ \text{over } s\text{-sparse } z \in X.$$

Moreover, for any s-sparse $z \in E$ norm $\|\cdot\|$ satisfies $\|z\| \le \sqrt{s}\|z\|_2$.

In what follows we say that x_s is a "sparsification of x."

Examples

1. "Vanilla" sparsity: in this case $E = \mathbb{R}^n$ with the standard inner product, and $\|\cdot\| = \|\cdot\|_1$. Assumption S2 clearly holds, e.g., when X is orthosymmetric, e.g., a ball of ℓ_p-norm on $\mathbb{R}^n$, $1 \le p \le \infty$.

2. Group sparsity... $\|x\| = \sum_{k=1}^K \|x_k\|_2$—block ℓ_1/ℓ_2-norm.

3. Low rank sparsity structure... $\|x\| = \sum_{i=1}^q \sigma_i(x)$ is the nuclear norm, $\sigma_1(x) \ge \sigma_2(x) \ge ... \ge \sigma_q(x)$ are singular values of x.

Proximal setup

• Let $\vartheta : E \to \mathbb{R}$ be a continuously differentiable convex function which is strongly convex w.r.t. norm $\|\cdot\|$, i.e.,

$$\langle \nabla\vartheta(x) - \nabla\vartheta(x'), x - x' \rangle \geq \|x - x'\|^2, \quad \forall x, x' \in E,$$

we assume that $\vartheta(x) \geq \vartheta(0) = 0$. We say that Θ is the constant of quadratic growth of $\vartheta(\cdot)$ if

$$\forall x \in E \; \vartheta(x) \leq \Theta\|x\|^2.$$

• If, in addition, Θ is "not too large," and for any $x \in X$, $a \in E$ and $\beta > 0$ a high accuracy solution to the minimization problem

$$\min_{z \in X} \{ \langle a, z \rangle + \beta\vartheta(z - x) \}$$

can be easily computed we say that distance-generating function (d.-g.f.) ϑ is "prox-friendly."

Stochastic Mirror Descent (SMD)

For $x, x_0 \in X$, $u \in E$, and $\beta > 0$ consider the proximal mapping

$$\mathrm{Prox}_\beta(u, x; x_0) := \operatorname*{argmin}_{z \in X} \{\langle u - \beta[\langle \nabla\vartheta(x - x_0), z\rangle - \vartheta(z - x_0)]\},$$

and for $i = 1, 2, \ldots$, consider Stochastic Mirror Descent recursion

$$x_i = \mathrm{Prox}_{\beta_{i-1}}(\nabla G(x_{i-1}, \omega_i), x_{i-1}; x_0), \quad x_0 \in X,$$

Here $\beta_i > 0$, $i = 0, 1, \ldots$, is a stepsize parameter, and $\omega_1, \omega_2, \ldots$ are independent identically distributed (i.i.d.) realizations of random variable ω, corresponding to the oracle queries at each step of the algorithm.

The approximate solution to problem (SO) after N iterations is defined as weighted average

$$\widehat{x}_N = \left[\sum_{i=1}^N \beta_{i-1}^{-1}\right]^{-1} \sum_{i=1}^N \beta_{i-1}^{-1} x_i.$$

Proposition 1 *Suppose that SMD algorithm is applied to problem (SO). We assume that Assumption S1 holds and that initial condition* $x_0 \in X$ *is independent of* ω_i, $i = 1,2,...$ *and such that*

$$\mathbf{E}\{\|x_0 - x_*\|^2\} \leq R^2;$$

we use constant stepsizes

$$\beta_i \equiv \beta \geq 2\varkappa\nu, \quad i = 1,2,...,m.$$

Then approximate solution

$$\widehat{x}_m = \frac{1}{m}\sum_{i=1}^{m} x_i$$

after m *steps of the algorithm satisfies*

$$\mathbf{E}\{g(\widehat{x}_m)\} - g_* \leq \frac{2R^2\Theta\beta}{m} + \underbrace{\frac{R^2\varkappa\nu^2}{\beta m}}_{I^{(1)}} + \underbrace{\frac{2\varkappa' \varsigma_*^2}{\beta}}_{I^{(2)}}.$$

Algorithm 1 [SMD-SR]

Parameters: R (initial error bound), $\bar{s}$ (upper bound for s), $\underline{\kappa}$, ...

1. Preliminary phase

Initialization: Set $y_0 = x_0$, $R_0 = R$, $\beta_0 = 2\varkappa\nu$, and $m_0 \asymp \frac{\bar{s}}{\underline{\kappa}}\Theta\varkappa\nu$.

Put

$$\overline{K} \asymp \ln_2\left(\frac{\underline{\kappa}}{\bar{s}}\frac{R_0^2\nu\varkappa}{\varsigma_*^2\varkappa'}\right)$$

and run

$$K = \min\left\{\left\lfloor\frac{N}{m_0}\right\rfloor, \overline{K}\right\}$$

stages of the preliminary phase.

Stage $k = 1,...,K$: Compute approximate solution $\widehat{x}_{m_0}(y_{k-1},\beta_0)$ after m_0 iterations of the SMD algorithm with constant stepsize parameter β_0, corresponding to the initial condition $x_0 = y_{k-1}$. Then define y_k as "s-sparsification" of $\widehat{x}_{m_0}(y_{k-1},\beta_0)$, i.e., $y_k = \mathrm{sparse}(\widehat{x}_{m_0}(y_{k-1},\beta_0))$.

Output: define $\widehat{y}^{(1)} = y_K$ and $\widehat{x}^{(1)} = \widehat{x}_{m_0}(y_{K-1},\beta)$ as approximate solutions at the end of the phase.

2. Set $M = N - m_0\overline{K}$ and

$$m_k \asymp \frac{\bar{s}\Theta\nu\varkappa}{\underline{\kappa}}2^k, \quad k = 1, 2, \ldots$$

If $m_1 > M$ terminate and output $\widehat{y}_N = \widehat{y}^{(1)}$ and $\widehat{x}_N = \widehat{x}^{(1)}$ as approximate solutions by the procedure; otherwise, continue with stages of the asymptotic phase.

Asymptotic phase

Initialization: Set

$$K' = \max\left\{k : \sum_{i=1}^{k} m_i \leq M\right\},$$

$y'_0 = \widehat{y}^{(1)}$, and $\beta_k = 2^k\nu\varkappa$, $k = 1, \ldots, K'$.

Stage $k = 1, \ldots, K'$: Compute $\widehat{x}_{m_k}(y'_{k-1}, \beta_k)$; same as above, define $y'_k = \mathrm{sparse}(\widehat{x}_{m_k}(y'_{k-1}, \beta_k))$.

Output: After K' stages, output $\widehat{y}_N = y'_{K'}$ and $\widehat{x}_N = \widehat{x}_{m_{K'}}(y'_{K'-1}, \beta_{K'})$.

What is going on?

• During the preliminary stage k, assume that $\mathbf{E}\{\|y_{k-1} - x_*\|^2\} \le R_k^2$ and $I_k^{(1)} \ge I_k^{(2)}$ in the bound of Proposition 1:

$$\mathbf{E}\{g(\widehat{x}_m)\} - g_* \le \frac{2R_k^2\Theta\beta}{m} + \underbrace{\frac{2R_k^2\Theta\varkappa\nu^2}{2\beta m}}_{I_k^{(1)}} + \underbrace{\frac{2\varkappa'\varsigma_*^2}{\beta}}_{I_k^{(2)}}$$

$\Rightarrow$ when choosing $\beta = 2\nu\kappa$ we get after $m_0 \asymp \Theta\varkappa\nu\bar{s}/\underline{\kappa}$ iterations

$$\mathbf{E}\{g(\widehat{x}_{m_0})\} - g_* \lesssim \frac{R_k^2\Theta\varkappa\nu}{m_0} \lesssim \frac{R_k^2\underline{\kappa}}{\bar{s}}.$$

• Due to quadratic lower bounding, $\|\widehat{x}_t - x_*\|_2^2$ decreases by factor $O(1/\bar{s})$, and because for $y_k = \mathrm{sparse}(\widehat{x}_{m_0})$

$$\|y_k - x_*\| \le \sqrt{2s}\|y_k - x_*\|_2^2 \le 2\sqrt{2s}\|\widehat{x}_{m_0} - x_*\|_2^2,$$

so that $\|y_k - x_*\|$ decreases by a constant factor and becomes $\le R_{k+1} = R_k/2$.

• As a result, after k preliminary stages of the algorithm,

$$\mathbf{E}\{\|y_k - x_*\|^2\} \quad \le \quad 2s\mathbf{E}\{\|y_k - x_*\|_2^2\} \lesssim 2^{-k}R^2 + \frac{\varsigma_*^2\bar{s}\varkappa'}{\underline{\kappa}\nu\varkappa} \lesssim \frac{\bar{s}}{\underline{\kappa}}\frac{\varsigma_*^2\varkappa'}{\nu\varkappa}.$$

after $k = \overline{K}$ preliminary stages.

- During asymptotic stage k, $I_2 \geq I_1$. When $\|y'_{k-1} - x_*\| \leq R_k$, the choice

$$\beta_k = \frac{\varsigma_*}{R_k}\sqrt{\frac{\varkappa m}{\Theta}}$$

results in

$$\mathbf{E}\{g(\widehat{x}_{m_k})\} - g_* \lesssim R_k \varsigma_* \sqrt{\frac{\Theta\varkappa}{m_k}},$$

$\Rightarrow$ after

$$m_k \asymp \frac{s^2\varsigma_*^2\Theta\varkappa}{\underline{\kappa}^2 R^2}$$

iterations

$$g(\widehat{x}_{m_k}) - g_* \lesssim \frac{R_k^2\underline{\kappa}}{\bar{s}}$$

and, by quadratic lower bound,

$$\|y_k - x_*\|^2 \leq 8s\|\widehat{x}_{m_k} - x_*\|_2^2 \leq \frac{16s}{\underline{\kappa}}[g(\widehat{x}_t) - g_*] \lesssim R_k^2/4 = R_{k+1}^2.$$

Main result Let $|\cdot|$ stand for $\|\cdot\|_2$- or $\|\cdot\|$-norm.

We define

- Recovery risk: maximal over $x_* \in X$ expected squared error

$$\mathrm{Risk}_{|\cdot|}(\widehat{x}|X) = \sup_{x_* \in X} \left(\mathbf{E}\{|\widehat{x} - x_*|^2\}\right)^{1/2}$$

- Prediction risk: maximal over $x_* \in X$ expected suboptimality

$$\mathrm{Risk}_g(\widehat{x}|X) = \sup_{x_* \in X} \mathbf{E}\{g(\widehat{x})\} - g_*.$$

Theorem 1 *Suppose that $N \geq m_0$, so at least one preliminary stage of Algorithm 1 is completed. Then approximate solutions $\widehat{x}_N$ and $\widehat{y}_N$ produced by the algorithm satisfy*

$$\mathrm{Risk}_g(\widehat{x}_N|X) \leq \frac{\underline{\kappa}R^2}{\bar{s}} \exp\left\{-\frac{cN\underline{\kappa}}{\Theta\varkappa\bar{s}\nu}\right\} + C\frac{\varsigma_*^2\bar{s}\varkappa'\Theta}{\underline{\kappa}N},$$

and

$$\begin{aligned}\mathrm{Risk}_{\|\cdot\|}(\widehat{y}_N|X) &\leq \sqrt{2s}\mathrm{Risk}_{\|\cdot\|_2}(\widehat{y}_N|X) \leq \sqrt{8s}\mathrm{Risk}_{\|\cdot\|_2}(\widehat{x}_N|X)\\ &\lesssim R\exp\left\{-\frac{cN\underline{\kappa}}{\Theta\varkappa\bar{s}\nu}\right\} + \frac{\varsigma_*\bar{s}}{\underline{\kappa}}\sqrt{\frac{\Theta\varkappa'}{N}}.\end{aligned}$$

Application to sparse linear regression

Consider the problem of recovery of a sparse signal $x_* \in \mathbb{R}^n$, $n \geq 3$, from independent and identically distributed observations

$$\eta_i = \phi_i^T x_* + \sigma\xi_i, \quad i = 1, 2, ..., N,$$

with ϕ_i and ξ_i mutually independent and such that $\mathbf{E}\{\phi_i\phi_i^T\} = \Sigma$, $\kappa_\Sigma I \preceq \Sigma$, and $\|\Sigma\|_\infty \leq v$,[1)] with known $\kappa_\Sigma > 0$ and v; we also assume that $\mathbf{E}\{\xi_i\} = 0$ and $\mathbf{E}\{\xi_i^2\} \leq 1$.

We are about to apply Stochastic Optimization to the problem

$$\min_{x \in X} \left\{ g(x) = \tfrac{1}{2}\mathbf{E}\{ \underbrace{(\eta - \phi^T x)^2}_{=:G(x,\omega=[\phi,\eta])} \} \right\}. \tag{SR}$$

We set $\|\cdot\| = \|\cdot\|_1$ with $\|\cdot\|_* = \|\cdot\|_\infty$, and we use "$\ell_1$-proximal setup" of the SMD-SR algorithm with quadratically growing for $n > 2$ distance-generating function

$$\vartheta(x) = \tfrac{1}{2}e\ln(n)\, n^{(p-1)(2-p)/p}\|x\|_p^2, \quad p = 1 + \frac{1}{\ln n},$$

the corresponding Θ satisfying $\Theta \leq \frac{1}{2}e^2 \ln n$.

[1)] For matrix Q we denote $\|Q\|_\infty = \max_{ij} |[Q]_{ij}|$.

Proposition 2 *Suppose that*

$$\mathbf{E}\{\|\phi\phi^T(x-x_*)\|_\infty^2\} \lesssim \varkappa\nu(x-x_*)^T\Sigma(x-x_*)$$

and that sample size N satisfy

$$N \geq m_0 \asymp \frac{\bar{s}}{\kappa_\Sigma}\nu\varkappa\ln[n],$$

so at least one preliminary stage of Algorithm 1 is completed.

Then approximate solutions $\widehat{x}_N$ and $\widehat{y}_N$ produced by the algorithm satisfy

$$\begin{aligned}
\mathrm{Risk}_{\|\cdot\|}(\widehat{y}_N|X) &\leq 2\sqrt{2s}\mathrm{Risk}_{\|\cdot\|_2}(\widehat{x}_N|X) \lesssim R\exp\left\{-\frac{cN\kappa_\Sigma}{\varkappa\bar{s}\nu\ln n}\right\} + \frac{\sigma\bar{s}}{\kappa_\Sigma}\sqrt{\frac{\nu\ln n}{N}} \\
\mathrm{Risk}_g(\widehat{x}_N|X) &\lesssim \frac{\kappa_\Sigma R^2}{\bar{s}}\exp\left\{-\frac{cN\kappa_\Sigma}{\varkappa\bar{s}\nu\ln n}\right\} + \frac{\nu\sigma^2\bar{s}\varkappa'\ln n}{\kappa_\Sigma N}.
\end{aligned}$$

• Remark: apart from positive definiteness of Σ, assumptions about regressor model essentially resume to

$$\mathbf{E}\{\|\phi\phi^T z\|_\infty^2\} \lesssim \nu \|\Sigma^{1/2} z\|_2^2 \ \ \forall z \in \mathbb{R}^n. \qquad (\Sigma_1)$$

Bound (Σ_1) holds in a variety of situations, e.g., it suffices that

$$\mathbf{E}\{(\phi^T z)^4\}^{1/2} \lesssim \mathbf{E}\{(\phi^T z)^2\} \ \ \forall z \in \mathbb{R}^n.$$

In particular, it holds in the case of

- bounded regressors such that $\|\phi_i\|_\infty$ a.s.
- sub-Gaussian regressors $\phi_i \sim \mathcal{SG}(0, S)$, i.e., meaning that

  $$\mathbf{E}\{e^{u^T\phi}\} \leq \exp\left\{\tfrac{1}{2} u^T S u\right\} \ \text{ for all } u \in \mathbb{R}^n.$$

  (Σ_1) holds when S is "similar" to the covariance matrix Σ of ϕ, i.e. $S \preceq \mu\Sigma$.
- scale mixtures $\phi \sim \sqrt{Z}\eta$, random variable $Z > 0$ a.s. with $\mathbf{E}\{Z^2\} < \infty$ and $\eta \in \mathbb{R}^n$ with $\mathbf{E}\{\eta\eta^T\} = \Sigma_0$ satisfies (Σ_1) and is independent of Z.

  E.g., $\phi \sim t_n(q, \Sigma_0)$ (n-dimensional Student distribution with q d.f.) with $q > 4$.
- ...

Low rank matrix recovery

• Let E be the space of real $p \times q$ matrices equipped with the Frobenius scalar product

$$\langle a, b \rangle = \mathrm{Tr}\left(a^T b\right).$$

Consider the problem of recovery of a $p \times q$ matrix x_*, from i.i.d. observations

$$\eta_i = \langle \phi_i, x_* \rangle + \sigma \xi_i, \quad i = 1, 2, ..., N,$$

with random regressors $\phi_i \in \mathbb{R}^{p \times q}$ having covariance operator

$$\Sigma(x) = \mathbf{E}\{\phi \langle \phi, x \rangle\},$$

$\xi_i \in \mathbb{R}$ independent of ϕ_i with $\mathbf{E}\{\xi_i\} = 0$ and $\mathbf{E}\{\xi_i^2\} \leq 1$.

• We put $\|a\|_2 = \langle a, a \rangle^{1/2}$, $\|\cdot\|$ is the nuclear norm $\|x\| = \|\sigma(x)\|_1$ where $\sigma(x)$ is the singular spectrum of x, and $\|y\|_* = \|\sigma(y)\|_\infty$ (the spectral norm).

• We suppose that

$$\kappa_\Sigma \|z\|_2^2 \leq \langle z, \Sigma(z) \rangle \leq v \|z\|_2^2 \ \ \forall z \in \mathbb{R}^{p \times q},$$

with known $\kappa_\Sigma > 0$ and v, and denote $\|z\|_\Sigma = \sqrt{\langle z, \Sigma(z) \rangle} = \sqrt{\mathbf{E}\{\langle \phi, z \rangle^2\}}$.

• We assume that x_* is of rank $s \leq \bar{s} \leq q$ and that we are given $R < \infty$ and $x_0 \in X$ satisfying $\|x_* - x_0\| \leq R$.

Consider the Stochastic Optimization problem

$$\min_{x\in X}\Big\{g(x)=\tfrac{1}{2}\mathbf{E}\{\underbrace{(\eta-\langle\phi,x\rangle)^2}_{=:G(x,\omega=[\phi,\eta])}\}=\tfrac{1}{2}\mathbf{E}\{(\sigma\xi+\langle\phi,x_*-x\rangle)^2\}=\tfrac{1}{2}(\|x-x_*\|_\Sigma^2+\sigma^2)\Big\}. \qquad \text{(LRR)}$$

• We are to solve (LRR) utilizing SMD-SR algorithm in the proximal setup associated with quadratically growing for $q\geq 2$ distance-generating function

$$\vartheta(x)=2e\ln(2q)\left[\sum_{j=1}^{q}\sigma_j^{1+r}(x)\right]^{\frac{2}{1+r}},\quad r=\big(12\ln[2q]\big)^{-1}.$$

with $\Theta\leq C\ln[2q]$.

• We suppose that regressors $\phi_i\in\mathbf{R}^{p\times q}$ are drawn from a *sub-Gaussian ensemble*, $\phi_i\sim\mathcal{SG}(0,S)$, with sub-Gaussian operator S, i.e.,

$$\mathbf{E}\{e^{\langle x,\phi\rangle}\}\leq e^{\langle x,S(x)\rangle/2}\ \ \forall x\in\mathbb{R}^{p\times q}$$

with linear positive definite $S(\cdot)$ which is similar to $\Sigma(\cdot)$, i.e., $S\preceq\mu\Sigma$ for some $\mu<\infty$.

Proposition 3 *In the just described situation, let the sample size* N *satisfy*

$$N \asymp \left[\frac{\mu^2 v(p+q)\bar{s}\ln q}{\kappa_\Sigma}\right],$$

so that at least one preliminary stage of Algorithm 1 is completed.

Then there is an absolute $c > 0$ *such that approximate solutions* $\widehat{x}_N$ *and* $\widehat{y}_N$ *produced by the algorithm satisfy*

$$\begin{aligned}
\mathrm{Risk}_{\|\cdot\|}(\widehat{y}_N|X) &\leq 2\sqrt{2s}\mathrm{Risk}_{\|\cdot\|_2}(\widehat{x}_N|X) \lesssim R\exp\left\{-\frac{cN\kappa_\Sigma}{\mu^2 v(p+q)\bar{s}\ln q}\right\} + \frac{\sigma\bar{s}}{\kappa_\Sigma}\sqrt{\frac{\mu v(p+q)\ln q}{N}}, \\
\mathrm{Risk}_g(\widehat{x}_N|X) &\lesssim \frac{\kappa_\Sigma R^2}{\bar{s}}\exp\left\{-\frac{cN\kappa_\Sigma}{\mu^2 v(p+q)\bar{s}\ln q}\right\} + \frac{\sigma^2\mu v(p+q)\bar{s}\ln q}{\kappa_\Sigma N}.
\end{aligned}$$

Minibatch implementation

One may use minibatches to save on prox-mapping computations when implementing the asymptotic phase of the algorithm. It amounts to replace gradient observation $\nabla G(x,\omega)$ with averages

$$\overline{\nabla G}(x,\omega^J) = \frac{1}{J}\sum_{j=1}^{J}\nabla G(x,\omega_j).$$

If choosing, at the k-th asymptotic stage,

$$J_k = 2^k\chi, \quad \beta_k = \beta_0 \asymp \varkappa\nu, \quad m_k = m_0 \asymp \frac{\bar{s}}{\underline{\kappa}}\Theta\varkappa\nu$$

the method only performs m_0 prox computations per stage.

- For instance, in sparse linear regression problem, when setting

$$\chi \asymp \ln n,$$

the error bounds for algorithm with minibatches coincide with those of Propositions 2 and 3 up to a logarithmic in n factors.

How it works

• Consider sparse linear regression with i.i.d. random (ϕ_i, ξ_i). In our experiments, Σ is diagonal with entries $\Sigma_{11} \leq \Sigma_{22} \leq \cdots \leq \Sigma_{nn}$ evenly spaced over $[\kappa_\Sigma, \nu]$, parameters (κ_Σ, ν) being specific for each experiment.

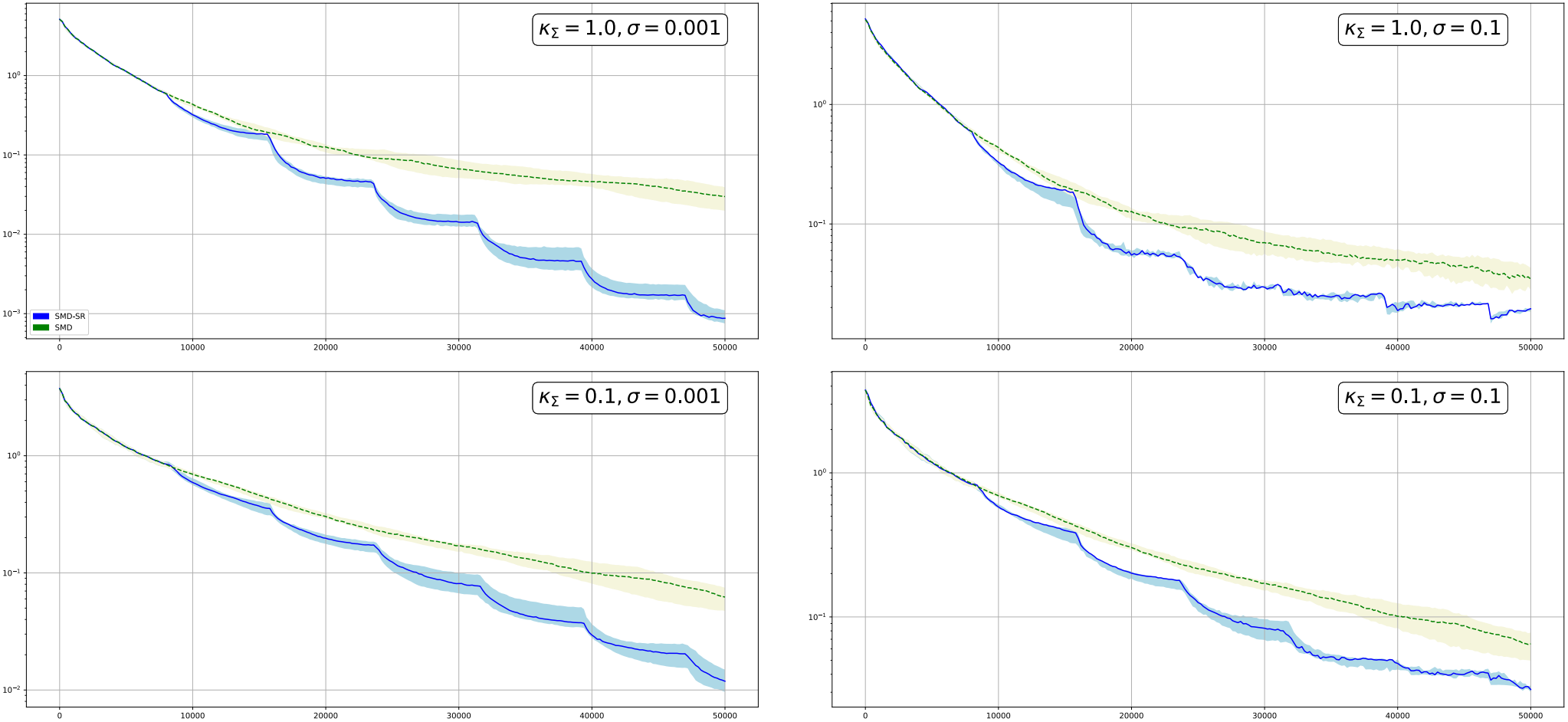
Indices of nonvanishing components of the optimal solution x_* are evenly spaced in $[1, n]$ with the non-zero entries sampled from $\mathcal{N}(0, 1)$.

• When solving (SR), we compare the performance of the SMD-SR to that of the "vanilla" non-Euclidean SMD and the Coordinate Descent algorithm (CDA) of the Python package `sklearn` solving Lasso problem

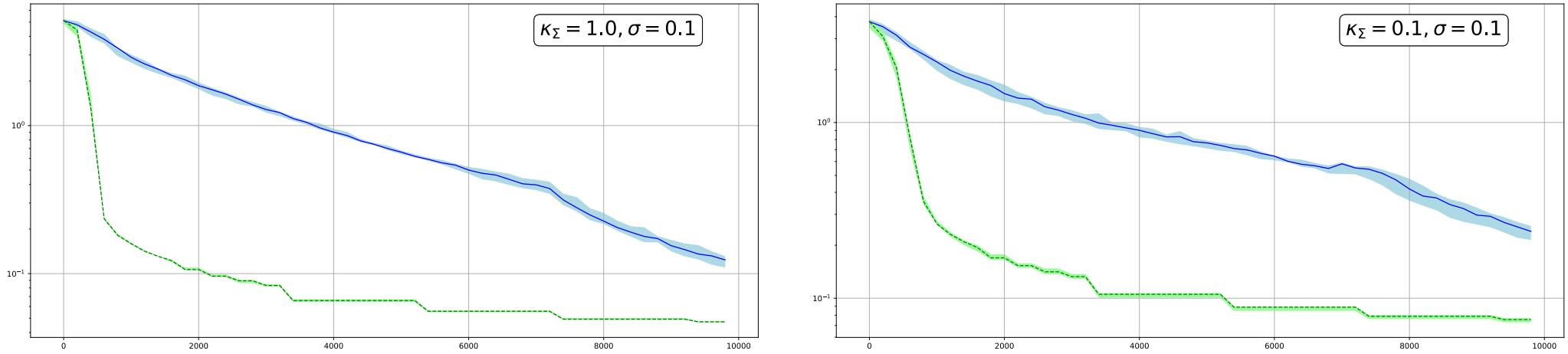
$$\min_{x \in \mathbb{R}^n} \left\{ \frac{1}{2N} \sum_{i=1}^{N} [\eta_i - \phi_i^T x]^2 + \kappa \|x\|_1 \right\}$$

with $\kappa = 2\sigma \sqrt{\frac{2 \ln n}{N}}$.

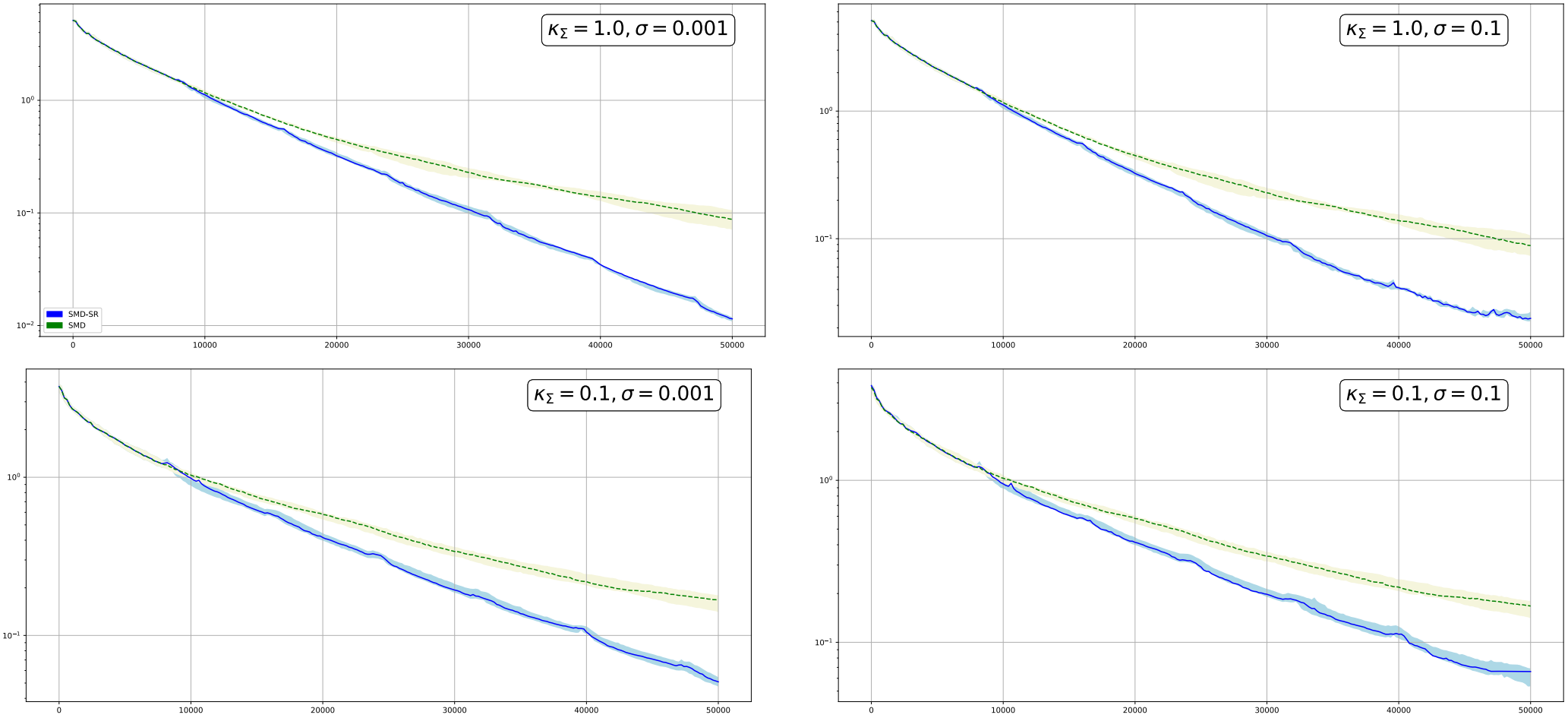
Experimental results



Comparison of SMD-SR (solid line) and SMD (dashed line) in the Gaussian setting; $(n, s) = (100\,000, 50)$.



Comparison of SMD-SR (solid line) and Lasso by CDA (dashed line) in the Gaussian setting; $(n, s) = (50\,000, 50)$.



Comparison of SMD-SR (solid line) and SMD (dashed line) in the case of Student t_4 regressors and noise distribution; $(n, s) = (100000, 50)$.

Some extensions

- Enhancing reliability of solutions utilizing Median-of-Means approach.

Suppose that available sample of length N can be split into L independent samples of length $M \asymp N/L$.

We may run Algorithm 1 on each subsample thus getting L independent recoveries $\widehat{x}_M^{(1)}, ..., \widehat{x}_M^{(L)}$, then compute an "enhanced solution" as a geometric median of $\widehat{x}_M^{(1)}, ..., \widehat{x}_M^{(L)}$,

$$\widehat{x}_{N,1-\epsilon} \in \operatorname*{Argmin}_x \sum_{\ell=1}^{L} \|x - \widehat{x}_M^{(\ell)}\|_2,$$

and then set $\widehat{y}_{N,1-\epsilon} = \mathrm{sparse}(\widehat{x}_{N,1-\epsilon})$.

- Reliable solution aggregation.

Assume that two independent observation samples of lengths N and $K \asymp N$ are available.

Same as above, we may use the first sample to compute L independent approximate SMD-SR solutions $\widehat{x}_M^{(\ell)}$, $\ell = 1, ..., L$, $M \asymp N/L$.

Then we can "aggregate" $\widehat{x}_M^{(1)}, ..., \widehat{x}_M^{(L)}$—select the best of them in terms of the objective value $g(\widehat{x}_M^{(\ell)})$ by computing reliable estimations of differences $g(\widehat{x}_M^{(i)}) - g(\widehat{x}_M^{(j)})$ using observations of the second sample.

Consider

- ϵ-risk of recovery

$$\mathrm{Risk}_{|\cdot|,\epsilon}(\widehat{x}|X) = \inf\left\{r : \sup_{x_*\in X} \mathrm{Prob}\{|\widehat{x} - x_*| \geq r\} \leq \epsilon\right\}$$

  where $|\cdot|$ stands for $\|\cdot\|_2$ or $\|\cdot\|$,

- ϵ-prediction risk

$$\mathrm{Risk}_{g,\epsilon}(\widehat{x}|X) = \inf\left\{r : \sup_{x_*\in X} \mathrm{Prob}\{g(\widehat{x}) - g_* \geq r\} \leq \epsilon\right\}.$$

Theorem 2 *Let $\epsilon \in (0, \frac{1}{4}]$, and let $\widehat{x}_{N,1-\epsilon}$ (resp., $\widehat{y}_{N,1-\epsilon}$) be a reliable solution by aggregating $L \asymp \alpha \ln[1/\epsilon]$ independent approximate solutions $\widehat{x}_M^{(1)}, ..., \widehat{x}_M^{(L)}$ by Algorithm 1. When $N \geq Lm_0$ we have*

$$\begin{aligned}
\mathrm{Risk}_{g,\epsilon}(\overline{x}_{N,1-\epsilon}|X) &\lesssim \frac{\kappa R^2}{\bar{s}} \exp\left\{-\frac{cN\underline{\kappa}}{\varkappa\bar{s}\nu\Theta\ln[1/\epsilon]}\right\} + \frac{\varsigma_*^2\bar{s}\varkappa'\Theta\ln[1/\epsilon]}{\underline{\kappa}N}, \\
\mathrm{Risk}_{\|\cdot\|,\epsilon}(\overline{y}_{N,1-\epsilon}|X) &\leq \sqrt{2s}\mathrm{Risk}_{\|\cdot\|_2,\epsilon}(\overline{y}_{N,1-\epsilon}|X) \lesssim R\exp\left\{-\frac{cN\underline{\kappa}}{\varkappa\bar{s}\nu\Theta\ln[1/\epsilon]}\right\} + \frac{\varsigma_*\bar{s}}{\underline{\kappa}}\sqrt{\frac{\varkappa'\Theta\ln[1/\epsilon]}{N}}.
\end{aligned}$$

- Penalized algorithm—condition of quadratic minoration of g for the restricted minoration condition

$$g(x) - g(x_*) \geq \mu\langle x - x_*, \Sigma(x - x_*)\rangle, \quad \forall x \in X,$$

where $\Sigma \succeq 0$ is such that for some $\chi < 1/2$, all $z \in E$ and all s-sparsifications z_s of z it holds

$$\|z_s\| \leq \lambda\sqrt{s}|z|_\Sigma + \chi\|z\|.$$

In this case, a "properly adjusted" implementation of the SMD algorithm for composite optimization solving a sequence of problems

$$\min_{x \in X} g(x) + \kappa_k\|x\| \qquad (\mathsf{P}_k)$$

attains the bounds analogous to those of Theorem 3.

- Stochastic Variational Inequalities with sparse solutions can be addressed when utilizing Mirror version of Popov's extragradient algorithm *Popov '80.*

- ...